EXPERIMENTAL PHILOSOPHY ON FREE WILL AND DETERMINISM

By

Robin Aldridge-Sutton

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Master of Arts

in Philosophy

Victoria University of Wellington

2011

Abstract

In the debate over whether free will is compatible with determinism most philosophers on both sides think that folk intuitions are a constraint on philosophical theorising. Most also think that the same criteria are required for free will in all cases. But recent attempts to empirically study folk intuitions about free will appear to show that these two positions cannot be jointly maintained. That is because folk intuitions about free will appear to represent compatibilist and incompatibilist criteria for free will in different cases. In response to this some philosophers have run new studies to demonstrate problems with older ones and undermine their results. One such study has been claimed to show that some participants mistakenly inferred that an agent's mental states have no effect on their actions given determinism.

In this thesis I argue that the questions about causation that were used in this study were too ambiguous to show this. My central point is that when considering the causal history of an action we tend to privilege the earlier causes over later ones. When participants responded that an agent's mental states have no effect on their actions they may have meant that there were earlier conditions that caused the agent's mental states which then caused their actions. This would show that the participants had made the correct inference that given determinism the causal histories of an agent's actions extend back in time to events outside of the agent's mind. Thus the problem of the apparent mismatch between folk intuitions and philosophical theories of free will remains. I also suggest that the results of these studies also appear to demonstrate a greater level of disagreement among folk intuitions about the same cases than philosophers seem to expect. This raises questions about whether the same theory of free will has to apply to everyone, and if so, whether folk intuitions support any such theory.

In chapter one I begin by briefly describing the free will debate and the role that folk intuitions usually play within it. I describe some debate over the usefulness of folk intuitions in philosophy, and make some small contributions on behalf of their usefulness. I describe and defend the recent movement towards attempting to empirically study folk intuitions on philosophical issues.

In chapter two I describe the empirical studies that seem to show that folk intuitions about free will represent compatibilist criteria in some cases and incompatibilist criteria in others. I suggest that they also seem to show that when considering the same case some folk have compatibilist intuitions and some have incompatibilist intuitions. I raise some questions about the implications of rejecting the assumption, which most philosophers seem to make, that people generally have the same intuitions about the same cases.

In chapter three I present the study that is claimed to show that participants in the earlier studies were confused, and thus that they did not really have incompatibilist intuitions. I present my arguments against it and conclude that for now the original interpretation of those studies stands, along with the problems it raises.

Acknowledgements

I am grateful to Ken Perszyk, Beth Romeril, Joanna Collins, and the postgraduate students at the philosophy department at Victoria University of Wellington.

Contents

<u>Introduction</u>	8
Chapter 1: <u>The Free Will Debate and Appeals to Intuitions</u>	13
1. <u>The Free Will Debate</u>	14
2. <u>Appeals to (Folk) Intuitions</u>	16
3. <u>Arguments Against Appealing to Folk Intuitions</u>	23
4. <u>Experimental Philosophy</u>	31
Chapter 2: <u>Experimental Philosophy on the Compatibility Question</u>	39
1. <u>Nahmias, Morris, Nadelhoffer, and Turner</u>	41
2. <u>Nichols and Knobe</u>	49
3. <u>Doris, Knobe, and Woolfolk</u>	61
4. <u>Variation in Folk Intuitions</u>	69
Chapter 3: <u>The Bypass Error Hypothesis</u>	77
1. <u>Nelkin</u>	78
2. <u>Nahmias and Murray</u>	85
3. <u>Problems with the Experiment</u>	97
4. <u>The Ambiguity Thesis</u>	108
<u>Conclusion</u>	111
<u>Bibliography</u>	115

Introduction

The most common issue in the free will debate has been whether free will (moral responsibility) is compatible with determinism. Compatibilists believe that free will is compatible with determinism, and incompatibilists believe that free will is incompatible with determinism. Both sides often appeal to intuitions in support of their views. Both sides think that folk (non-philosophers') intuitions or beliefs are a constraint on philosophical theorising, though they acknowledge that sometimes folk intuitions will need to be revised or rejected. (This has been called the thesis of conservatism (Doris, et al., 2007:184).) While they disagree about what the criteria for the attribution of free will or moral responsibility are, they both assume that there are exceptionlessly relevant criteria. That is, the same criteria apply in every case or context. (This has been called the thesis of invariantism (Doris, et al., 2007:184).)

Recently, some philosophers – known as experimental philosophers – have begun to employ the methods of empirical psychology to directly study the nature of folk intuitions in a range of philosophical areas. These areas include epistemology, philosophy of science, ethics, and philosophy of mind. In recent years experimental philosophers have increasingly focused their attention on the free will debate. Early studies appear to have found that the intuitive judgments of the folk are split, with a majority supporting compatibilism, contrary to what incompatibilists have led us to believe. However, later studies appear to have found that the majority of folk intuitions support incompatibilism in different contexts.

It has been claimed, for example by Doris, Knobe, and Woolfolk (2007), and Knobe and Doris (2010), that these findings have significant implications for the free will debate. For while philosophers tend to think that folk intuitions are a constraint on philosophical theorising (conservatism), they also hold that there is a single set of criteria for attributing moral responsibility that should apply in all contexts (invariantism). It appears that one or other of these positions must change in response to the apparent fact that folk intuitions support compatibilism and

incompatibilism in different contexts. Doris, Knobe, and Woolfolk recommend that conservatism is kept but invariantism rejected.

In response to this challenge some philosophers have proposed that the folk intuitions may not be as they appear, i.e. that they do not support compatibilism and incompatibilism in different contexts. For example, Nelkin (2007) argues that the apparent incompatibilist intuitions found in the above studies may be based on a misinterpretation of determinism. If the folk properly understood what determinism means, then they would have compatibilist intuitions in all contexts. Nahmias and Murray (2010) ran experiments to test this hypothesis. They argue that the majority of folk who appear to be incompatibilists mistakenly interpret determinism to imply that a person's mental states have no effect on their actions. Those who do not misunderstand determinism tend to be compatibilists.

In this thesis I will argue for two main points. The first is that even within the same context there was more disagreement among folk intuitions than is commonly assumed in the debate. In contrast to Doris, Knobe, and Woolfolk, who argue for conservative variantism, I will only argue that there is a good case in favour of variation in folk intuitions, both between different contexts as they claim, and between different folk in the same contexts, which has not often been discussed. This raises interesting questions about the applicability of the debate to different people, and which intuitions, if any, are fit to support philosophical theories. The second main point is that there are several problems with Nahmias and Murray's study that is supposed to show that incompatibilist intuitions are based on a misinterpretation of determinism. The most important problem is that the questions designed to assess how people understand the effects of mental states in a deterministic scenario are ambiguous. Some participants plausibly interpreted them in a different way than was intended. On this interpretation their responses would not indicate that they had made the error that was hypothesized, but that they had understood the implications of determinism correctly. Thus the incompatibilist intuitions are not undermined, and the questions raised by the original results remain to be considered.

I will begin in chapter one by describing and supporting the role that appeals to folk intuitions have had in the free will debate and in philosophy more generally. I will respond to several arguments against appealing to folk intuitions, and argue that while it may be difficult to discover reliable folk intuitions, it is worthwhile for such an intractable debate. I will then describe and advocate the use of experimental philosophy to explore the nature of folk intuitions and the processes that generate them. By employing the methods of modern empirical psychology it avoids many biases that philosophical and classroom discussion risks in that study.

In chapter two I will describe two experimental studies that seem to have shown that folk intuitions support both sides of the free will debate. I will not be attempting an exhaustive review of the experimental philosophy literature on free will. In the first study, Nahmias, Morris, Nadelhoffer, and Turner (2005), it appears that when presented with a concrete description of a particular agent performing a particular action in a deterministic world, most people judged that they were praiseworthy or blameworthy for doing so, and that they did so of their own free will. This was taken to suggest that the majority of the folk are pretheoretic or natural compatibilists, contrary to the “received view”. In the second study, Nichols and Knobe (2007), this result was apparently replicated; but when participants were presented with an abstract description of a deterministic world alone, and asked whether agents in general could be responsible in such a world, most people judged that they could not be fully morally responsible. This was taken to suggest that the folk have both compatibilist and incompatibilist intuitions in different contexts.

I will then describe the argument that has been put forward by Doris, Knobe, and Woolfolk (2007), that this variation in folk intuitions requires a change in the standard philosophical approach to the debate. It is that either philosophers must retract their assumption that folk belief is a constraint on philosophical theorising, or they must reject their assumption that the criteria for responsibility attribution are exceptionlessly relevant. As they say, “The reason for this is straightforward: *folk morality is not invariantist*” (2007:185).

I will then discuss the idea that the data that seems to show that folk morality is not invariantist also seems to show that within the same context folk intuitions are

different for different people. Regardless of whether the majority appear to support compatibilism or incompatibilism, there is almost always a significant minority who have the opposite intuition. In some cases intuitions approach an even split. I will briefly describe another experimental study (Feltz, et al., 2008) which suggests that different people have different intuitions about free will, based on their personality traits. This kind of variation is also a problem for most philosophers who assume that folk intuitions about free will exhibit wide-spread agreement. As this agreement seems to underlie the ability of intuitions to support philosophical theories, it is not clear whose intuitions, if anybody's, are fit to do so. Perhaps different theories apply to different people. Perhaps we should prioritise the intuitions of the majority. Perhaps no one's intuitions are common enough. Philosophers must either consider these issues or explain how folk intuitions about free will may not be as they appear in these studies.

Finally, in chapter three I will present an argument for the claim that the apparent variation in folk intuitions is the product of a misinterpretation on the part of some of the participants. It has been proposed that descriptions of determinism tend to be misinterpreted, for example that determinism implies that people would be coerced, or that an agent's mental states would have no effect on what they do. (This echoes a long-standing compatibilist complaint, that incompatibilists confuse causation with compulsion.) It is an attempt to explain away apparent incompatibilist intuitions as being due to an error. The study that sought to test this hypothesis (Nahmias and Murray, 2010) replicated two earlier experiments but also asked participants whether an agent's mental states had no effect on what they ended up being caused to do. There was a strong correlation between responses that agents' mental states have no effect on what they do and that they are not responsible for what they do. I will criticise this experiment. I will argue that questions about the efficacy of an agent's mental states are ambiguous. The responses that were interpreted as saying that an agent's mental states have no effect on what they do under determinism may have meant that both an agent's mental states and what they do are determined by prior conditions outside of their control. Thus they may not indicate an error. Thus the study does not show that apparent variation in folk intuitions about free will is based on a misunderstanding of determinism. (This reply echoes the long-standing incompatibilist denial of confusing causation with compulsion; the problem, they

say, if determinism is true is not that the prior conditions or laws of nature control us but that the source of control is not in or up to us.) Further research will be required to settle whether this is what those responses meant. In the mean time the philosophical commitments to there being widely-shared folk intuitions, and to both conservatism and invariantism, are at risk of being untenable.

Chapter 1: The Free Will Debate and Appeals to Intuitions

In section one I will briefly introduce the free will debate and outline the main positions within it.

In section two I will describe the wide-spread practice in philosophy of appealing to (folk) intuitions, in general and particularly in the free will debate. Intuitions are not of course always reliable, they can be in conflict with each other, and they can be based on a misunderstanding of the issues. But by surveying intuitions in a range of cases philosophers can identify dominant and well-grounded intuitions which are most likely to indicate the ordinary criteria for responsibility attribution. These criteria are important because they are the subject of strong beliefs and feelings, which are widely shared and influence much of our behaviour.

In section three I will consider some arguments for the claim that philosophers should not appeal to intuitions, especially folk intuitions. What the evidentiary role, if any, of intuitions in philosophy is, is a deep meta-philosophical question. My aim in this thesis is not to settle this question. It may well be that philosophers sometimes misuse or abuse intuitions, but I will argue that folk intuitions are relevant to the free will debate.

Finally, in section four I will argue that the best way to study folk intuitions is that of a recent movement known as experimental philosophy. I will begin by describing this movement which employs the tools of empirical psychology to study the way the folk think about philosophical questions. I will argue that it is well motivated because philosophers on both sides of the free will debate often claim that most folk intuitions accord with their views, and this disagreement speaks against their having reliable knowledge about them. I will then reply to some arguments that experimental philosophy is unnecessary because philosophers can learn enough about folk intuitions in other ways. I will then argue that while experimental philosophy is still in its infancy, and its early results are only tentative, it has the potential to distinguish reliable intuitions from superficial ones. The free will debate is particularly intractable and involves an explicit disagreement about the nature of

folk intuitions, and experimental philosophy is precisely what is required to resolve this dispute.

1. The Free Will Debate

The most common or dominant question in the free will debate, since at least the 17th century, is the compatibility question, which is whether free will (moral responsibility) is compatible with determinism. That is, whether it is possible for free will to exist if determinism is true. If free will is incompatible with determinism, the falsity of determinism is not sufficient for the existence of free will. There remains the important question of whether free will is compatible with indeterminism.

The term “free will” can be used to mean different things, but perhaps the most dominant meaning in the literature, and the one I will be considering, is the set of conditions - whatever they are – that are required for someone to be morally responsible for an action (Vargas, 2007:128). There are also different senses of “responsibility” – e.g. causal, legal, and moral. In addition, sometimes people seem to have in mind a ‘forward-looking’ sense (deterrence); at other times a ‘backward-looking’ sense (punishment). The term “determinism” can also mean different things and determinism comes in different forms.¹ Following Kane (2005:5-6) the core idea in all its forms is that an event is determined just in case there are earlier conditions whose occurrence is logically sufficient for the occurrence of the event. That is, necessarily given those earlier conditions, the determined event occurs or will occur. The kind of determinism that is most discussed these days is causal

¹ Given that there are several possible meanings of the terms “free will”, “responsibility” and “determinism”, this means that there is no single compatibility question. This may be important when it comes to assessing the studies discussed in chapters two and three, because it leaves open the possibility that the folk may be incompatibilists about some senses of “free will” (“responsibility”) and “determinism” but not about others.

determinism. There are different ways to express the thesis of causal determinism. On one reading, it is the thesis that for every event, except the first (if there is one), the past and the laws of nature entail the occurrence of that event. Indeterminism, in this context, is the thesis that causal determinism is false. To say that our choices or actions are causally undetermined is to say that they can occur otherwise given exactly the same past and laws of nature. Compatibilism is the view that it is possible for free will to exist if determinism is true², and incompatibilism is the view that it is not.³

The significance of the compatibility question lies in the implication that if incompatibilism is true, then the ordinary belief that we are free and responsible for at least some of our actions can only be true if determinism is false. If determinism is true, we are never free or responsible. Some incompatibilist philosophers think that determinism is true, and so they reject our ordinary belief that we are free or

² Fischer's view is known as semi-compatibilism: "The doctrine of semi-compatibilism is the claim that causal determinism is compatible with moral responsibility, quite apart from whether causal determinism rules out the kind of freedom that involves access to alternative possibilities." (2007:56)

³ Incompatibilists think that one or both of the following is (are) required for free will: (1) the existence of alternative possibilities – one must be able to do otherwise; (2) sourcehood – i.e. the agent must be the (ultimate) source of her actions. Some compatibilists (viz. semi-compatibilists) and some libertarians and other incompatibilists (viz. those persuaded by Frankfurt (1969)) deny the first requirement. Traditional compatibilists agree that the first condition is required, but they disagree with incompatibilists over what's required to satisfy it. That is, they disagree over the sense of "could have done otherwise". Traditional compatibilists (unlike incompatibilists) typically give a conditional analysis: I would have done otherwise if I had different beliefs, desires, etc.

responsible for our actions. This view is known as *hard determinism*. It is very rarely supported, probably because it would seem to have extreme implications for how we feel about ourselves and how we treat each other.⁴ Other incompatibilist philosophers think that we are sometimes free or responsible, and so they deny determinism. These are known as *libertarians*.

One might wonder why determinism is a threat, given that most experts in quantum physics believe that determinism is false for the behavior of at least some very tiny particles. One reason is that quantum indeterminacy is generally thought to cancel out at the much greater scale involved in human behavior. Thus while universal determinism is probably false, it does not obviously mean that *human behavior* is undetermined.⁵ In addition, even if human behavior is not determined that is hardly sufficient to ensure the existence of free will or moral responsibility. Suppose your choice or action is the result of some undetermined quantum event in your brain. This would seem to make it due to chance, in which case it is hard to see how you would have sufficient control to be responsible for it. This problem plagues libertarian accounts of any kind, some more than others.⁶

2. *Appeals to (Folk) Intuitions*

⁴ For example Smilansky (2000) thinks that if we do not foster the illusion of free will it will have dire consequences for us and society. On the other hand, Honderich (1993) and Pereboom (2001) think that we can do well without the illusion of free will, though some important changes are needed. Pereboom's own position is called *hard incompatibilism*. It consists of the claims that free will is incompatible with determinism, and that absent agent-causation indeterministic causal histories are no less of a threat than deterministic causal histories.

⁵ See Kane (2005:7-10) for other reasons.

⁶ See for example O'Connor (1995).

In the free will debate, as in many areas of philosophy, intuitions are often used to support arguments. Intuitions are judgments of what seems to be true. They have often been thought to be the most basic premises that philosophers can begin with (Gutting, 1998:7). While the intuitiveness of general principles is also taken into account, the intuitions that are most commonly used in the free will debate, and elsewhere, are classificatory judgments about thought experiments (Nelkin, 2007:251). That is, judgments about whether a particular case, real or hypothetical, belongs to a certain class or category. (For example, an intuition that is the subject of much debate at the moment regards so called “Frankfurt cases” which I will return to shortly.) This kind of intuition (which is the kind I will be concerned with in the rest of this thesis) has been employed in philosophy in some sense since Socrates (Alexander and Weinberg, 2007:56). What they seem to do is show us the criteria by which we make classificatory judgments. By analysing the cases in which we attribute or refrain from attributing a certain property, such as responsibility for an action, we can extrapolate the conditions that govern those attributions. In the free will debate we are arguably concerned with ordinary thinking about responsibility. When most folk have the same intuitions about whether agents in certain thought experiments are responsible or not for their actions, we can make inferences about the conditions that ordinarily govern the attribution of responsibility.

It is a standard assumption in much of philosophy that conformity with common sense is a constraint on philosophical theorising. There is a presumption in favour of theories that fit best with, or explain, wide-spread intuitions. Being counter-intuitive is typically thought to be a cost, though of course it may not be decisive.

Philosophers often give competing analyses of concepts, for example what is required for free will or moral responsibility. Asking others i.e. folk, is one way to try to adjudicate between these analyses.

No one thinks that folk intuitions settle a debate on their own, as if it follows that whatever is intuitive is true. Philosophical theories should also systematize them, revise inconsistencies amongst them and with scientific facts, or explain how they

are not as they seem (Nahmias and Murray, 2010:4).⁷ Revising inconsistencies amongst our intuitions is part of the process of seeking a reflective equilibrium. Nelkin (2007:251) notes that philosophers who rely on appeals to intuitions typically aim to find a reflective equilibrium (see Rawls 1971) between intuitions and principles, acknowledging that there may be some conflicts in our intuitions. Finding a reflective equilibrium roughly involves identifying such conflicts and weighing up the intuitiveness of each side. In this way philosophers can infer the conditions which would tend to be prioritised in situations that make such conflicts salient.

Intuitions may also be explained away because they may be based on a misunderstanding or on an invalid inference. Philosophers try to make thought experiments as clear and unambiguous as possible to avoid this problem. Once reliable intuitions are discovered philosophers try to systematise them by finding necessary and sufficient conditions for the attribution of the concepts they involve. Folk intuitions provide evidence about ordinary thinking which philosophers try to describe as simply and thoroughly as they can. Nahmias, Morris, Nadelhoffer, and Turner (henceforth NMNT) think that when an intuition is common it shifts the burden of proof in a debate (2005:564).⁸

⁷ Page numbers for citations of Nahmias and Murray's paper are taken from the manuscript that was available prior to the recent publication cited. It is available at <http://www2.gsu.edu/~phlean/papers.html>

⁸ It has been argued that incompatibilism has a greater burden of proof due to the relative complexity of, for example, Libertarian accounts, as opposed to compatibilist accounts, of free will. If incompatibilism is not the intuitive view then it is not clear why we should accept the more complex view (Nahmias et al., 2006:30-33).

There is a wide-spread intuition that we are free and responsible for at least some of our actions. This seems to be fundamentally connected to our view of ourselves as persons, and our place in the universe.⁹

There are two main ways that philosophers appeal to intuitions in the free will literature. One is to appeal to whether compatibilism or, more commonly, incompatibilism is the ordinary or intuitive view. The other is to appeal to whether certain arguments for incompatibilism, or principles involved in them, are intuitive.

Regarding the first, the most common claim is that incompatibilism is the ordinary or intuitive view. For example:

Most ordinary people start out as natural incompatibilists Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers (Kane, 1999:218).

We come to the table, nearly all of us, as pretheoretic incompatibilists (Ekstrom, 2002:310).

[Libertarian free will is] the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess (Strawson, 1986:30); [it is] in our nature to take determinism to pose a serious problem for our notions of responsibility and freedom (Strawson, 1986:89).

Most of us start off by making an important assumption about freedom. Our freedom of action, we naturally tend to assume, must be incompatible with our actions being determined (Pink, 2004:12).

People naturally assume that they have libertarian free will, and it has formed the basis of most of the ethical teaching of the Western religions and of major ethical systems such as Kant's (Smilansky, 2003:259).

⁹ See for example Fischer and Ravizza (1998) chapter 1, and Kane (1996) chapter 6.

When ordinary people come to consciously recognize and understand that some action is contingent upon circumstances in an agent's past that are beyond that agent's control, they quickly lose a propensity to impute moral responsibility to the agent for that action (Cover and O'Leary-Hawthorn, 1996:50).

Beginning students typically recoil at the compatibilist response to the problem of moral responsibility (Pereboom, 2001:xvi).

Does freedom of choice have this implication [that causal determinism must be false]? It seems so to the typical undergraduate on first encountering the question (O'Connor, 2002:4).

The really interesting and controversial question is about the *conditions* of moral responsibility, and in particular the question whether freedom of a contra-causal kind is among these conditions. The answer of the common man to the latter question is that it most certainly *is* among the conditions (Campbell, 1951:451).

The following are examples in which compatibilists claim that their view is intuitive:

[The "Reason View", that the ability to choose the 'True' and the 'Good' is what is required for free will,] seems to accord with and account for the whole set of our intuitions about responsibility (Wolf, 1990:89).

[Compatibilism is] the default position . . . not only true, but the only position rationally available to impartial observers (Lycan, 2003:107).

Incompatibilists often appeal to intuitions in support of their principles and arguments. One of the most famous arguments for incompatibilism is van Inwagen's (1983) Consequence Argument. The argument relies on an inference rule, what he calls *rule Beta*, which is that if p implies q and no one has a choice about whether p implies q, and p is true, and no one has a choice about whether p is true, then q is true and no one has a choice about whether q is true. If Beta is valid then it looks as

though no one has a choice about anything if determinism is true, and so is never free, given that no one has a choice about the past or the laws of nature, or the fact that if determinism is true then the past and laws of nature entail everything that happens. This argument assumes that having alternative possibilities, or the ability to do otherwise, is required for freedom. Van Inwagen says:

I must confess that my belief in the validity of Beta has only two sources, one incommunicable and the other inconclusive. The former source is what philosophers are pleased to call “intuition”: when I carefully consider Beta it seems to be valid... The latter source is the fact that I can think of no instances of Beta that have, or could possibly have, true premises and a false conclusion (1983:97-98).

Another example of an incompatibilist appealing to intuitions in support of an argument is Pereboom regarding his Four Case Argument for the incompatibility of physical determinism and free will (2007:93-98).¹⁰ Physical determinism is the conjunction of causal determinism and physicalism, the view that everything in the world is physical. In the Four Case Argument Pereboom describes four cases in which Professor Plum decides to murder Ms White, and does so, for some personal advantage, and as a result of a deterministic causal process. He specifies that in each case the prominent compatibilist conditions for moral responsibility are met. In the first case Plum is created and manipulated by neuroscientists in a way that determines that he will murder Ms White. In the second case Plum is only created by neuroscientists in a way that determines that he will murder Ms White. In the third case Plum is subject to rigorous training practices that determine that he will murder Ms White. In the fourth case Plum is raised in normal circumstances but physical determinism is true and so he is causally determined to murder Ms White. Pereboom thinks that it is clear and intuitive that the agent is not free or responsible for the murder in the first three cases despite satisfying prominent compatibilist conditions for being free and responsible. His challenge to compatibilists is to give a reason that Plum is responsible in the fourth case despite not being responsible in the preceding cases. He claims that the non-responsibility of the agent in at least one of

¹⁰ This is slightly different from Pereboom (2001) where he does not include physicalism in this argument.

those cases generalises to the fourth case, because he thinks that the best explanation of the non-responsibility in the first three cases is that the murder results from a deterministic process that leads back to factors beyond Plum's control. Because Plum is determined to act in the same way in the final case he is also not responsible there. Thus he argues that moral responsibility is incompatible with determinism.

Compatibilists often appeal to intuitions in proposing counter-examples to incompatibilist principles or arguments. The most famous examples involve "Frankfurt-style cases" (following Frankfurt, 1969). In these cases an agent is described as unknowingly having a device in her brain that can predict and control her behaviour. The device can predict whether she will choose to do a particular thing (for example voting a certain way), and if she is going to choose to do it, it will intervene before her choice and cause her to choose not to do it. In fact she chooses not to do it without the device intervening. Upon considering this thought experiment many philosophers, including some "source incompatibilists" (such as Pereboom), have the intuition that she is responsible for her choice, because she did it for reasons of her own, even though she could not have chosen otherwise because the device would not have let her. This intuition is taken to show the falsity of the Principle of Alternative Possibilities (PAP), i.e. that an agent is responsible for doing A only if she could have done otherwise. The falsity of PAP is then used (by some compatibilists) to undermine the Consequence Argument for the incompatibility of free will or responsibility with determinism. Even if determinism implies that you cannot do otherwise, the ability to do otherwise is not required for freedom/responsibility. Dennett (1984:558) claims that when ordinary people assign moral responsibility, "it simply does not matter at all . . . whether the agent in question could have done otherwise in the circumstances."

Fischer and Ravizza (1998) give what they think are intuitive counter-examples to what they call the Transfer of Non-Responsibility Principle, which is roughly equivalent to van Inwagen's rule Beta. One such example is "Erosion" (1998:157) in which Betty plants her explosives, causing an avalanche that crushes the enemy. Unbeknownst to her there is an eroding glacier. If she had not placed the explosives, there would have been an avalanche that would have crushed the enemy. The intuition is that Betty is morally responsible for placing the explosives, and crushing

the enemy, even though there were conditions for which Betty was not responsible, and which were sufficient to cause the avalanche and crush the enemy. In other words, the following inference is invalid: there is an eroding glacier and no one is even partly responsible for there being an eroding glacier; if there is an eroding glacier then there is an avalanche that crushes the enemy and no one is even partly responsible for this fact, therefore there is an avalanche that crushes the enemy and no one is even partly responsible for this fact.

While incompatibilists and compatibilists disagree about either what conditions are required for free will or moral responsibility, or whether they can be satisfied under determinism, they agree that intuitions are important here.

3. Arguments Against Appealing to Folk Intuitions

Having given a brief account of the way incompatibilists and compatibilists appeal to (folk) intuitions in the free will debate, I will now consider some general challenges to this practice. Some philosophers, such as Jonathan Weinberg (2007), strongly doubt whether intuitions, folk or philosophers', are useful evidence at all. Weinberg thinks that the lack of a detailed account of the origins of intuitions means that there is no reason to think that they can tell us anything important about the world. While I will not be addressing this concern in depth, it is worth pointing out some reasons for thinking that it is worth studying intuitions about free will.

First of all, it is thought that almost everyone has strong intuitions about the conditions under which someone is responsible for what they do. We often feel quite sure about whether a person in a particular case is responsible, and this is thought to extend to cases involving determinism. Secondly, it is thought that there is wide-spread agreement about these issues. While philosophers hotly dispute whether free will (moral responsibility) is compatible with determinism, they tend to think (as we have seen) that the folk are on their side, and that some of their colleagues are confused. Thirdly, those intuitions are thought to inform a great deal of our behaviour and emotions. Whether or not someone is thought to be responsible for something influences how we feel about them, and how we act towards them. This includes assessments of our own responsibility. Reactive attitudes like anger,

gratitude, guilt, and pride are thought to be tied up with responsibility (Strawson, 1962). Praise and blame, and punishment and reward are thought to be more or much less appropriate depending on it. Intuitions about responsibility are supposed to tell us something important about the world because they are generally thought to be strongly felt, widely shared, and of great practical importance.¹¹ Whether or not this is true is a deep question that applies to the use of intuitions in many areas of philosophy. If appeals to intuitions are illegitimate, then much of philosophy is in trouble. While I will not attempt to grapple with this issue directly, it is clear (as we have seen) that both incompatibilists and compatibilists regularly appeal to (folk) intuitions in support of their views.

It is perhaps more common to challenge the use of folk intuitions in philosophy because they are *folk* intuitions rather than because they are intuitions. Alexander and Weinberg discuss several arguments against the use of folk intuitions specifically (2007:57-60). One view is that it is individual philosophers' own intuitions that are the right evidence. They think that this view is probably unpopular but that there are a few high-profile endorsements of it. For example, they cite Kornblith as noting that "David Lewis writes as if he conceives of himself as doing just this when he writes that he discovered his own intuitions about counterfactuals and then attempted to construct a theory which successfully captured those intuitions" (2007:76). But Alexander and Weinberg point out that this view does not seem to make sense of how most philosophers conceive of their practice. Usually philosophers express their claims about intuitions using the impersonal locution "it is intuitive that", the second-person-involving "it should be apparent to the reader that..." and especially the first-person-plural "our intuition is that..." (when they are not just expressing the views of several authors). They also think that appealing only to one's own intuitions would not make sense of the argumentative context of philosophy. We are not just describing our own intellectual lives but trying to prove theses. The epistemic value of intuitions must come either from their being widely shared or from the person who has them having some recognised authority in the relevant area. Considering only one person's intuitions is not sufficient to establish that they have either of these qualities. If some philosophers

¹¹ Compare for example Kane (2007:5-7) and Fischer (2007:46-47).

have been successful by consulting their own intuitions perhaps it indicates that they are experts in their areas, or that their intuitions are widely shared, but this could not have been known on the basis of their intuitions alone (Alexander and Weinberg, 2007:58).

Alexander and Weinberg think that it is a better challenge to the use of folk intuitions to claim that only philosophers' intuitions should be used in philosophy (Alexander and Weinberg, 2007:58). They think that *intuition elitism*, as they call this view, at least makes sense of philosophical practice as an argumentative enterprise. They consider several arguments for it. One that they think is common, though not in print, is based on the fact that philosophical claims often involve technical concepts. It has been suggested that only philosophers' intuitions can support such claims because only they reflect an adequate understanding of the concepts involved. I will call this "the argument from technical concepts".

Alexander and Weinberg think that it fails because the concerns that originally lead us to philosophical investigation have to do with ordinary concepts, and that if we only consider technical concepts we risk losing touch with those original concerns (2007:58).

I agree with the premise of the argument from technical concepts that a person's intuitions are only as good as their understanding of the concepts involved. Though I do not think that it is obvious that all philosophical investigation is motivated by issues to do with ordinary concepts, the free will debate certainly is. I agree with several philosophers who have claimed that folk intuitions may be irrelevant to some inherently abstract philosophical issues, but that they are relevant to philosophical issues concerned with everyday moral views and practices, including ascriptions of free will and moral responsibility (Nahmias and Murray, 2010:4, Nadelhoffer and Nahmias, 2007:130, Knobe and Nichols, 2008:8).

The argument from technical concepts raises an important point. The concept of determinism is certainly technical enough to make one wonder whether folk intuitions can reflect an adequate understanding of it. It is often claimed that determinism is not a thesis that most folk believe, and there is some empirical support for this. In one of the experimental studies that I will discuss in chapter two,

participants were presented with a deterministic universe, described so as to try to make the determinism accessible to them, and an indeterministic universe where people's choices were not determined by prior conditions. A large majority judged that the indeterministic universe was more similar to our own (Nichols and Knobe, 2007:669). So it seems likely that the idea of determinism is not an ordinary concept, and thus that the folk may not generally understand it well. Philosophers use intuitions as evidence of the nature of free will, rather than of determinism. But you still need to understand determinism in order to have useful intuitions about its implications for free will. If a person does not understand the implications of determinism, then asking them whether their intuition is that free will is compatible with determinism is not going to get us anywhere. So I think that the argument from technical concepts can raise a valid concern for the usefulness of folk intuitions to the free will debate, but there are reasons to think it does not undermine their use completely.

As I have noted, the intuitions used in the free will debate are judgments about whether an agent is, or could be, morally responsible for an action in a certain situation. This is supposed to be more useful than simply judging the compatibility question directly. Part of the motivation for this technique is plausibly that it allows for features of determinism to be expressed in a way that is easy to understand, and thus makes it easier to judge their consequences. One way this can work is by replacing a lengthy description of the concept of determinism with a concrete example. Thus we do not have to understand determinism as an abstract concept but simply a particular scenario. This is precisely the technique used in the experimental studies which are the subject of chapters two and three. Constructing thought experiments is something of an art, and so whether the folk can understand determinism or not is a question that should be considered for particular thought experiments on a case by case basis. It requires empirical testing to answer with any confidence. Again, this is precisely what is being done in the experimental studies that this thesis discusses. They include questions, known as manipulation checks, designed specifically to test whether the participants understand the thought experiments properly. The majority of participants pass these checks, and those who do not are excluded or analysed separately. So, while the argument from technical concepts provides a good reason to be cautious, it does not seem to rule out the use

of folk intuitions in the free will debate. There remains a concern or worry here, viz. that despite the care being taken in the construction of these thought experiments to avoid the reliance on an abstract thesis of determinism which could easily be misunderstood, too much work is being done by the wording of the particular examples. The difficulty is coming up with wording that does not beg the question of whether free will is compatible with determinism.¹²

Another argument for intuition elitism that Alexander and Weinberg consider is the idea that if philosophers have a better record of epistemic success than do non-philosophers, this shows that their intuitions are more reliable (Alexander and Weinberg, 2007:59). But they think that most scientists are not philosophers and that it is doubtful that philosophers could demonstrate a greater level of predictive and explanatory success than they have. Also mathematicians have a fantastic record of answering open questions in their field so it is no more likely that philosophers could show that they are relatively better on this front either. And so it does not seem that philosophers are more epistemically successful than non-philosophers (2007:60). I think that, however successful philosophers may be in some areas, it would be hard to say that they have had much success in the free will debate. While their theories and arguments become more abundant and complex, it is not clear that the disagreement is coming closer to being resolved. It seems like their intuitions have led them into stalemates (Fischer, 1994:83-85).

Knobe and Nichols (2008:8) describe another common argument for intuition elitism. It is that since it would be absurd for physicists and biologists to survey folk intuitions about complex matters in physics and biology, the same is true of philosophy. While I agree that folk intuitions may not be useful in *every* area of philosophy, this does not mean that there are *no* areas in which they are useful. In addition, I think that in fact some scientists do study folk intuitions. Cognitive psychologists and linguists are two clear examples (for a review see Ramsey, 1998). Furthermore these are exactly the areas of science that are most similar to the free will debate. The subject of cognitive science is ordinary thinking. Both cognitive science and linguistics seek to understand the processes and information that

¹² See for example Sommers (2010:204).

underlie ordinary thinking, and which are not always directly consciously accessible. Thus they study ordinary people, and their judgments of concrete examples. Research in these areas is arguably no less successful than in other sciences for using this method. If philosophers should imitate the methods of scientists, they should imitate those who have the most similar goals to their own.

Nadelhoffer and Nahmias (2007:129) note the mounting evidence of heuristic biases, cognitive diversity, and situationism, in the thinking of experts as well as non-experts. Heuristic biases are tendencies to employ models which are quick, but sometimes give wrong answers, to think about certain problems. Cognitive diversity refers to the fact that different people think about things in different ways. Situationism is the influence of situational factors, such as the way a problem is framed or the problems that have been considered immediately prior to it, that influence the way we think. Given that these issues all apply to the thinking of experts as well as laypersons, Nadelhoffer and Nahmias (2007:129) think that a positive reason is required to expect philosophers' intuitions to be more reliable than the folk's.

Alexander and Weinberg (2007:59) note that it is common to claim that philosophers have a special competency for attending to the features of thought experiments that are relevant to philosophical claims. But while they accept that this would be a good reason to prioritise philosophical intuitions, they agree with Nadelhoffer and Nahmias that it requires an argument to support it. One possibility that they consider is that philosophers become competent because they spend more time thinking about thought experiments than the folk. But they do not find this convincing. They think that intuitions diverge among philosophers so that it is not obvious that reflection alone is an effective way to refine their intuitions. This seems to me to be the case in the free will debate where intuitions sometimes lead to dialectical stalemates, or arguments and counter-arguments becoming ever more numerous and complex without leading to more agreement (Fischer, 1994:83-85, Kane, 2007:13). It seems that the premises of some arguments are intuitive to some philosophers and not others, so that it is very difficult to resolve disagreement about them.

In contrast, Knobe (2006:82) thinks that it is a striking fact that the readers of analytic philosophy usually do share the intuitions that are appealed to in the philosophical literature. As an experimental philosopher perhaps he means that philosophical intuitions are relatively convergent, as compared to folk intuitions. This would support the argument that philosophers are relatively competent at judging thought experiments, based on the assumption above that convergence suggests progress, even if they are not perfect. But it is an empirical claim that requires evidence to assess, which Knobe does not provide. It is not clear whether philosophers' intuitions are more convergent than the folk's. But I think that regardless of this it is in itself quite plausible that spending time thinking about lots of different thought experiments and making judgments about them could make someone better at it. They would be more likely to interpret thought experiments in the ways that they are intended, and less likely to make invalid inferences about them. They are also more likely to be aware of conflicts in their intuitions between different thought experiments and weigh up which are stronger. Even if they are subject to heuristic biases, cognitive diversity, and situationism to the same extent as the folk, they may still be better in these ways.

Alexander and Weinberg (2007:59) think that even if the time philosophers spend thinking about thought experiments develops their proficiency, it is not clear that they spend significantly more time considering thought experiments than the folk. They think that it may be that philosophers spend more time considering what is to be made of the intuitions that they have, than on producing the right intuitions. This seems a bit hasty. They are right that it is not clear that philosophers spend most of their time producing intuitions about thought experiments rather than thinking about their intuitions themselves. But it still seems quite plausible that they are exposed to thought experiments much more often than the folk because that is an integral part of their training and occupation. It may be that some philosophers by deliberation or chance focus mostly on the implications of whatever intuitions are appealed to, but the opposite may also be true. Since exposure to thought experiments probably develops some relevant competence, it is likely that in general philosophers are to some extent better than the folk at producing reliable intuitions.

But folk intuitions are better than philosophical intuitions in other ways, as evidence of ordinary thinking. Philosophers are a demographically unrepresentative sample of the folk (Stich and Weinberg, 2001:642). People who study and work in analytic philosophy live overwhelmingly in western societies. They are generally white men. They are relatively very well-educated, and are very wealthy compared to most people in the world. It is at least a live question whether the folk in general have the same intuitions as this group of people. The best way to disprove this concern may be to do some experimental philosophy.

The folk are also in a more objective position than philosophers regarding what intuitions are common. Philosophers gain their qualifications and build their careers by defending philosophical theories, very often on the basis of intuitions. This experience seems likely to bias their intuitions, or the intuitions that they pay the most attention to, in favor of their own theories. This would mean that their theories would be driving their intuitions, rather than their intuitions indicating the nature of ordinary thinking. It is a common concern among philosophers that their intuitions may become contaminated with their theories in this way (Goldman and Pust, 1998:183, Nadelhoffer and Nahmias, 2007:125)¹³. Kauppinen (2007:96) notes that for intuitions to serve, as they are supposed to, as neutral data upon which competing conceptual analyses can be assessed, they must be the judgements of non-partisans to the debate. In this way folk intuitions can play a crucial role in resolving disagreements about ordinary concepts in philosophy. When such disagreements are not easily resolved in the course of philosophical discussion, we can look to folk intuitions to see what the relevant concepts really are.

This is precisely the situation in the free will debate. Philosophers strongly disagree about what the conditions required for free will and moral responsibility are, and whether they are compatible with determinism. This makes it implausible that philosophers' intuitions are a reliable source of information about it. It is reasonable to think that as a skewed sample of people, and as participants in the debate, they are

¹³ In the free will literature Fischer is particularly sensitive to this concern. See for example (2006:236 and 238).

the wrong people to ask. A natural step to take seems to be to explain the key point in the debate in simple and concrete terms to a sample of the folk and ask them what they think. It is likely to be more difficult to ensure that they understand the thought experiments and questions in the right way, and to identify and assess the relative strength of any conflicting intuitions they may have. But it is worth trying in order to resolve such an intractable disagreement. Until it is tried we have no good reason to think it cannot be done. Some philosophers have recently begun to try, and the movement has come to be known as *experimental philosophy*.

4. *Experimental Philosophy*

Experimental philosophy involves the application of modern methods of experimental psychology to the study of folk intuitions about philosophical thought experiments (Alexander and Weinberg, 2007:60). Experimental philosophers use controlled and systematic studies to explore the nature of intuitions and the cognitive processes that influence them. They present groups of non-philosophers with questionnaires including a description of some philosophically relevant scenario and questions about their classifications of it. They then statistically analyze the results, and examine the philosophical implications. Thus experimental philosophy is similar to traditional philosophy in considering intuitions about thought experiments. But experimental philosophers deny that we can know enough about them from armchair reflection. They consider claims about intuitions to deserve empirical testing.

Nadelhoffer and Nahmias (2007:123) point out that experimental philosophy composes a variety of projects with different interests, assumptions, and goals. They distinguish three different projects in experimental philosophy, all of which will be discussed in this thesis (2007:126-129). Many experimental philosophers simply test philosophers' claims about how common certain intuitions are, in order to assess the arguments based upon them. For example, Knobe (2003) tested whether most people have the intuition that the side effects of actions are brought about intentionally. Experimental philosophers follow traditional philosophy in taking intuitions provoked by thought experiments to provide interesting information that cannot always be discovered by direct reflection on philosophical questions. This

project is only in direct conflict with traditional philosophy in that it takes it to be necessary to study intuitions with the methods of experimental psychology.

Knobe makes the point that many of the papers in experimental philosophy are primarily concerned with discovering the internal psychological processes underlying people's intuitions (2006:90). This goes a step further than the project above by comparing sets of responses to different thought experiments to try and discover what factors influence people's judgments about them. Rather than just trying to accurately chart the spread of intuitions in a population, it prioritises the discovery of variables that have significant effects on their production. Knobe thinks that some of these experimental studies are interested in these processes for their own sake. Nadelhoffer and Nahmias describe this project similarly, but deny that it is undertaken for its own sake (2007:127). Rather, they think that it is used to evaluate philosophical theories on the basis of how the mind works. They give the example of work by Greene and others (for example, Greene, 2002, 2003) which concludes that moral judgment works on a sentimentalist, rather than a rationalist basis.

Other papers in experimental philosophy draw a third kind of conclusion. They seek to evaluate the processes underlying intuitions in terms of their capacity to produce intuitions that are reliable evidence. In this way they seek to restrict the set of intuitions that are used to support philosophical claims. This is very similar to the traditional consideration of possible error theories that seek to show that some intuitions are not good evidence. It can test whether the intuitions discovered are appropriate to support philosophical claims. It is thus directly relevant to the project of assessing philosophers' claims about how common intuitions are. It is the commonality of reliable intuitions that philosophers are concerned with.

Some experimental philosophers who try to assess the reliability of intuitions based on the psychological processes which underlie them draw quite radical conclusions. They think that both traditional philosophers and other experimental philosophers are mistaken in their assumption that intuitions in general have some evidentiary value

in philosophy.¹⁴ They think that even if we empirically discover exactly who has what intuitions they will not be fit to support philosophical claims. Generally it is because they are sensitive to irrelevant factors. Some of the factors suggested are cultural and educational background, and socio-economic status. Some experimental philosophers have claimed to have shown this for intuitions about several thought experiments in epistemology regarding whether someone really knows or only believes something (Nichols, et al., 2003, Weinberg, et al., 2001). As I mentioned in section two, one of the reasons philosophers take intuitions to be important is that they are supposed to be subject to wide-spread agreement. Some experimental philosophers claim to have shown that there is significant disagreement based on demographic differences. If true, this would seem to undermine one of the reasons to think that intuitions are important. As I mentioned in section two, I cannot discuss in detail such fundamental questions in this thesis, but it is a challenge for most philosophers to consider. It could turn out to be the case that differences in intuitions in the free will debate turn on irrelevant facts. I will briefly consider this possibility in chapter two. This is something that experimental philosophy should test for.

The main difference between experimental and traditional philosophy is the effort that is thought to be required to discover the intuitions that support philosophical claims. Many philosophers probably think that experimental philosophy is unnecessary. Some may think that only philosophers' intuitions matter, and that these are easily accessible. I have explained why I think folk intuitions matter. Philosophers who agree that folk intuitions can be of value to philosophy may think that philosophers know all they need to know about them already. In the free will debate philosophers often make claims (as we have seen) about the nature of folk intuitions, but they have not previously set out survey results to defend them. But the view that philosophers already know enough about folk intuitions seems implausible given that both compatibilists and incompatibilists often claim that their view is intuitive to the folk.

¹⁴ For a review and a sophisticated argument see Weinberg (2007).

Some philosophers argue that they can find out about folk intuitions more easily than by employing experimental philosophy. Kauppinen (2007:109-110) thinks that discussion with friends and family, other philosophers, and their students, is sufficient for philosophers to know what people find intuitive. But Nadelhoffer and Nahmias (2007:131) have pointed out that such dialogue will be influenced by where the participant grows up, their educational history, the variety of people they converse with most often, and the circumstances in which those conversations take place. These factors may lead people to have different ideas about what ordinary thinking is like. Furthermore, philosophical training may cause those ideas to become entrenched in the process of defending them against other philosophers. So discussion does not obviously solve either the problem of philosophers being a demographically unrepresentative sample of people, or of their theories contaminating their intuitions.

Frank Jackson has made a point that might seem apt to show that philosophers are at least able to resolve the threat of theory contamination without undertaking full-blown experimental studies (1998:36-37). He notes that they are continuously coming into contact with non-philosophers and engaging in philosophical discussion with them in their role as teachers. He thinks that informal polling of students in the class-room can give philosophers plenty of information about what non-philosophers think. While he does not make this point in the context of theory contamination it seems like a reason that some philosophers may think that experimental philosophy is unnecessary.

But class-room discussion as a method of discovering folk intuitions is actually very problematic (Stich, 2009:240). If philosophers' intuitions are different from the folk's then the way that they present philosophical problems is also likely to be biased. While philosophers try to present issues in an objective way it is far from clear how successful they are. It is not uncommon for philosophers to mention in discussion how difficult this can be for the free will debate in particular. This could mean that the students' reactions would not represent their views about the issues themselves but about a biased interpretation of them. This is exacerbated by the fact that in philosophy classes, especially in introductory ones, the primary goal is not to gauge students' intuitions but to educate them about the issues. The students are

trying to learn rather than to express and refine their own views. Again philosophers try to teach students to think for themselves, but this is often one of the hardest things for students to pick up when they are first introduced to philosophy. This could lead students to ignore their initial intuitions in favor of those that they perceive that they are supposed to have. They may think that their original intuitions are just mistakes based on a misunderstanding of the issues. In psychology there is plenty of evidence that subjects react in the ways that they think they are expected to. Features of an experiment which communicate such expectations are known as demand characteristics, and are avoided as a source of bias. As philosophical theories are complex, technical, and unfamiliar to the subjects, they require a lot of explanation. This explanation is likely to compound any demand characteristics that are present. So the way that information is communicated to philosophy students is very likely to bias their intuitions about philosophical problems.

Stich (2009:240) notes that there are also problems in the reporting and interpretation of students' in-class responses to thought experiments. In informal in-class polling students must express their views in a very public way. This has been shown in psychology to suppress dissenting opinions. Students who have intuitions that are different from the majority are less likely to express them, meaning that the strongest tendencies to respond in certain ways will be exaggerated. This also seems likely to cause many students to ignore their initial intuitions because they think that they are idiosyncratic. Thus this kind of polling may actually contribute to a process of indoctrination just as worrying for philosophy as theory contamination. A recent development is the use of "clicker" devices distributed to students in lectures that allow them to indicate their intuitions more privately. This may mitigate the tendency for minority responses to be suppressed. However, the results are often then displayed in the lecture, meaning that those who respond with the minority may still take their intuitions to be undermined to some extent upon learning this, especially if the lecturer is not sympathetic towards them.

In the interpretation of in-class polls one problem is known in psychology as experimenter bias. The fact that philosophers already have strong commitments to certain views being intuitive means that they are likely to misinterpret the reactions of students to accord with their expectations. They are likely to take students'

reactions to confirm what they already believe about whether something is intuitive, a form of confirmation bias. A simple example of this would be taking the number of people putting up their hands in a lecture theatre to be a greater majority than it is. Altogether these problems show that while the idea of informal polling to overcome theory contamination is in the right spirit, it is clearly flawed given some of the knowledge which underlies the empirical methodology that experimental philosophers employ.

Experimental philosophy employs empirical psychology to study folk intuitions in a rigorous way. By taking folk intuitions as its primary target it keeps them separate from the other processes involved in seeking reflective equilibrium. It eliminates philosophers from the sample of subjects studied thus making no assumptions about the relation between philosophers' and the folk's intuitions. By employing thought experiment style scenarios in questionnaires isolated from further discussion it is able to report precisely how the issues are presented to the subjects. Thus it can hold those presentations up to analysis and defend them from questions of bias. It can also include manipulation checks, questions that test for correct understanding of the scenarios, in the questionnaires. Thus subjects who do not understand the issues can be identified and their responses excluded from analysis, or otherwise reconsidered. Using questionnaires also separates the process of finding out what people think from that of educating them about major theories in the area. Thus demand characteristics can be reduced leaving subjects free to report their original intuitions about an issue rather than discounting them when they conflict with popular academic theories. They are also able to report their views privately, so that they are not given reason to think that others do not share their views, or required to publicly dissent from a perceived majority view. It avoids experimenter bias by giving relatively clear cut results which can be reported separately from their interpretation. The standardization of presentation and responses allows for replication of studies and statistical analysis. Statistically precise degrees of confidence in the ability of any effects found to be generalized can be calculated. Thus the cognitive processes that produce different and conflicting intuitions can be explored in great detail. By statistically testing the relatively subtle differences between response sets experimental philosophers can show significant variations that would be virtually invisible to traditional philosophy. Demographically representative samples of the

folk can be used. Most experimental philosophy so far has used university students, thus keeping many of the problems with sample bias that traditional philosophy has. But some has focused on cross-cultural research allowing for some of the cultural and economic biases to be tested. There are problems that apply to experimental philosophy also, such as that discussed, of ensuring that participants understand scenarios in the ways that they are intended.

One obvious concern about experimental philosophy is that while philosophers are experts at constructing and criticizing thought experiments they are not qualified in the psychological research methods involved. It is obvious, and acknowledged, that inter-disciplinary collaboration is required, and this is one thing that many experimental philosophy studies have and may still lack. But several have done this already, and it is quickly becoming common. Those studies that have not done this can still be analyzed as preliminary results, and that is generally how their authors have presented them. Their similarity to studies that do have qualified collaborators speaks in their favor.

Kauppinen (2007) and Sosa (2009) have both argued that experimental philosophy is actually insufficient to discover the right intuitions on which to base philosophical claims. Kauppinen argues that dialogue and reflection are required to find the robust kind of intuitions required for philosophical investigation (2007:105-107). He thinks that folk intuitions matter but only carefully formed ones rather than ‘surface intuitions’. Sosa has the related concern that some experimental philosophy thus far has not been able to show clearly that the subjects understand the questionnaire in the way that the experimenters believe (2009:107-109). Thus its results may represent quite different intuitions than they are claimed to.

These are both reasonable concerns, but they do not show that experimental philosophy is doomed to be a waste of time. What they show is that it is a difficult project that is still being tested out and developed. But I think that these concerns should be assessed on a case by case basis. If a survey appears to show that people have certain intuitions, then unless there are specific reasons to think it does not, we should think that it does. And if there are such reasons we should put them to the test empirically too. As will become clear in the discussion of the most recent

experimental results on the compatibility question, in chapter three, questionnaires can actually be made quite complex in terms of testing the ways that people understand them. But experimental philosophy is not necessarily tied to questionnaires. As Nadelhoffer and Nahmias point out it could easily move into behavioural studies and fMRI studies (2007:139), to correlate intuitions with behaviours and neural processes, or carefully standardised interviews (2007:131). In psychology qualitative and quantitative methods are often used to complement each other in the study of complex phenomena. Thus rather than writing experimental philosophy off, I think that Kauppinen and Sosa's criticism can help to develop it.

There have been a substantial number of experimental studies on the compatibility question in the last 10 years, and they are especially well motivated in this area. They concern a concept (moral responsibility) that the folk have strong beliefs about. The free will debate is widely acknowledged as particularly intractable among philosophers. It is subject to explicit disagreements about the nature of common folk intuitions. And while many philosophers have motivated their position in it by appealing to folk intuitions, they have not attempted to seriously test their claims. This is precisely what experimental philosophy seeks to do. Again, resolving this dispute would not thereby resolve the free will debate itself, as philosophers must still systematize, weigh up, and possibly revise those intuitions. But it is a crucial first step in this process.

Chapter 2: Experimental Philosophy on the Compatibility Question

In this chapter I will present the results of two experimental studies on folk intuitions about free will and argue that, unless some underlying error is found, they raise some problems for assumptions in the free will debate. In sections one and two I describe the studies and how they were originally interpreted. In the first study (Nahmias, et al., 2005) participants were presented with a description of a particular agent performing a particular action in a deterministic universe. This has come to be known as a ‘concrete’ scenario (Nichols and Knobe, 2007:669). The study apparently found that most of the participants judged that the agent was responsible for the action. It was interpreted as meaning that, contrary to the common claim described earlier, most people do not find incompatibilism intuitive.

In the second study (Nichols and Knobe, 2007) participants were presented with a description of a deterministic universe, and asked whether agents in general could be responsible for their actions. This has come to be known as an ‘abstract’ scenario (Nichols and Knobe, 2007:669). The study apparently found that most participants judged that agents could not be responsible for their actions in a deterministic universe. Different participants were then presented with a concrete deterministic scenario that either involved an act that was likely to arouse strong negative emotions, or an act that was not. It was apparently found that when the act aroused strong emotions most people judged that the agent was responsible for it, but when it did not most people judged that the agent was not responsible for it. This was interpreted as meaning that emotion plays an important role in compatibilist judgments. It was also taken to mean that it is likely, though not clear, that strong emotion causes people to respond as compatibilists when in fact they are not.

In section three I will describe an argument put forward by Doris, Knobe, and Woolfolk (2007) that these results raise a problem for standard philosophical assumptions in the free will debate. One of those assumptions is that folk intuitions form a constraint on theorising about free will (conservatism). The other assumption is that the correct theory should posit exceptionlessly relevant criteria for moral responsibility (invariantism). As folk intuitions appear to support compatibilism and incompatibilism in different contexts these assumptions do not seem to be consistent

with the data. Either conservatism or invariantism must be rejected unless the compatibilist or incompatibilist intuitions are based on an error. Doris, Knobe, and Woolfolk argue that it is not obvious that either the compatibilist or the incompatibilist intuitions are based on an error, and that invariantism should be rejected in favour of conservatism.

In section four I will present some considerations in favour of the view that the results of the studies described in sections one and two raise questions about standard assumptions in the free will debate. I will argue that, contrary to Doris, Knobe, and Woolfolk's view, it is not clear whether concreteness or emotion is the most plausible explanation of the differences in folk intuitions about the various scenarios. But I will also argue that it is not obvious that those differences are due to an error either way. Thus, while I will not follow Doris, et al., in arguing for rejecting either conservatism or invariantism specifically, I will support their case that the data raises a problem for philosophers who wish to maintain both views.

I will then distinguish another kind of variation in the results of the studies described in chapters one and two. Most of the discussion has focused on the fact that people appear to have different intuitions about different scenarios. I will call this 'intra-personal' variation, because it is supposed to be variation in intuitions within individual persons. But the results described also appear to show that different people have different intuitions about the same scenarios. I will call this 'inter-personal' variation. I will also briefly describe another experimental study that has been claimed to suggest that there is inter-personal variation in intuitions about free will based on the personality trait 'extroversion' (Feltz, et al., 2008). It appears that extroverts are much more likely to have compatibilist intuitions. I think that if there is inter-personal variation, that is not due to an error, it undermines the standard assumption in the debate that folk intuitions about free will are widely-shared. I will briefly consider some possible implications of rejecting this assumption. These include the possibility that the free will debate applies differently to different people, that the intuition of the majority should be prioritised, or that both or neither compatibilist nor incompatibilist intuitions are common enough to support their respective theories.

1. Nahmias, Morris, Nadelhoffer, and Turner

NMNT made a preliminary and exploratory attempt to use experimental philosophy to resolve the dispute about whether most of the folk find compatibilism or incompatibilism intuitive (Nahmias, et al., 2005). They ran the first rigorous experimental study of people's intuitive responses regarding the free will and responsibility of an agent in a deterministic scenario. The participants were undergraduate university students at Florida State who had not studied the free will problem. The first questionnaire they developed was based on the following scenario:

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 supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195.

Do you think that, when Jeremy robs the bank, he acts of his own free will?
(2005:566)

On the back participants answered a manipulation check, a question designed to ascertain whether they understood the scenario correctly. Those who failed were excluded from analysis. NMNT also excluded people who answered "I don't know" to the experimental question. They found that a statistically significant majority of the participants (76%, 16/21) thought that Jeremy robs the bank of his own free will (2005:566). They then repeated the experiment with different participants and a praiseworthy act, saving a child from a burning building, and a neutral act, going jogging. They did this to test whether a negative emotional response was priming

participants to hold the agent responsible, and that this, along with a tacit assumption that only an agent with free will can be responsible, was causing people to attribute free will. They found that the responses closely tracked those for the blameworthy act and were statistically significant, 68% (15/22) and 79% (15/21) saying that Jeremy acts of his own free will respectively (2007:567). So they think that a negative emotional response does not account for the results.¹⁵

They then repeated the blameworthy and praiseworthy cases with different participants again, but asked “Do you think that, when Jeremy robs the bank [saves the child], he is morally blameworthy [praiseworthy] for it?” Responses regarding moral blameworthiness and praiseworthiness also tracked those regarding free will, though with slightly greater majorities, and were statistically significant, 83% (15/18) and 88% (16/18) affirming the attributions respectively (2005:568). Participants were also invited to explain their answers, and NMNT note that some of them cited fledgling compatibilist or incompatibilist theories in doing so, though NMNT do not give any examples of these (2005:580). This seems to suggest that both compatibilism and incompatibilism have some roots in folk theories, given that the participants had never studied the free will problem.

NMNT also tested a different set of participants on the questions “Do you think that he could have chosen *not* to rob the bank [save the child/go jogging]?” (2005:568). That is because philosophers often hold that the ability to do otherwise is required for responsibility but is incompatible with determinism. Here they found that in the blameworthy case responses tracked those regarding free will and moral blameworthiness, with 67% (14/21) saying Jeremy could have chosen not to rob the bank, a statistically insignificant difference. But in the praiseworthy and neutral cases the majority of participants judged that Jeremy could *not* have chosen not to

¹⁵ The act of going jogging is unlikely to arouse any strong emotions in the participants at all, and yet received similar responses to the others. This will be important when it comes to discussing the intra-personal variation in the results, which I will do in section 4.

save the child or to go jogging, 62% (13/21) and 58% (8/14) respectively, which were both statistically significant differences.

They think that these intriguing results may suggest that people have the kind of intuitions predicted by arguments based on Frankfurt cases. That is that an agent may be responsible for an act despite not being able to do otherwise. They think this intuition may be more pronounced in praiseworthy and neutral cases, as has been proposed by Wolf (1980). They also suggest that in the blameworthy case people employ a conditional analysis of ability. That is a sense of ability in which one is able to act in a certain way if one would act that way if one wanted to (there are different variations in which different mental states are required). There are different conditional analyses of ability, but the idea behind them is roughly that someone is able to choose otherwise despite determinism being true because if they had wanted something else they would have chosen differently. I think that it is difficult to analyze these results because it is not clear what sense of ‘could have done otherwise’ participants were using, and NMNT did not attempt to find out. They seem to agree, noting that the complexity in judgments about the ability to choose otherwise shows that further research is required to understand them.

One might worry that the Laplacean notion of determinism involved in this scenario is not equivalent to, or as threatening to free will as, causal determinism as I have set it out. That is, the predictability of events based on the past and laws of nature may not be as important as events being causally determined by the past and laws of nature. In a later paper NMNT report the results of two experiments based on other descriptions of determinism to address the worry that determinism was not sufficiently salient in the earlier one (2006:37). One of the scenarios was the following:

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same conditions and the same laws of nature produce the exact same outcomes, so that every single time the universe is re-created, everything must happen the exact same way. For instance, in this universe a person named Jill decides to steal a necklace at a particular time, and every

time the universe is re-created, Jill decides to steal the necklace at that time.
(2006:38)

Subjects were also asked the manipulation check “According to the scenario, is it accurate to say that every time the universe is re-created, Jill makes the same decision?” (2006:38) and were excluded from analysis if they answered no. The results were similar to those of the previous experiment. 66% of participants judged that Jill acted of her own free will, and 77% judged that “it would be fair to hold her morally responsible (that is, blame her) for her decision to steal the necklace.” (sample sizes not reported)¹⁶ (2006:38).

The other scenario that NMNT employed was the following:

Imagine there is a world where the beliefs and values of every person are caused completely by the combination of one’s genes and one’s environment. For instance, one day in this world, two identical twins, named Fred and Barney, are born to a mother who puts them up for adoption. Fred is adopted by the Jerksens and Barney is adopted by the Kindersons. In Fred’s case, his genes and his upbringing by the selfish Jerkson family have caused him to value money above all else and to believe it is OK to acquire money however you can. In Barney’s case, his (identical) genes and his upbringing by the kindly Kinderson family have caused him to value honesty above all else and to believe one should always respect others’ property. Both Fred and Barney are intelligent individuals who are capable of deliberating about what they do.

One day Fred and Barney each happen to find a wallet containing \$1000 and the identification of the owner (neither man knows the owner). Each man is

¹⁶ The failure to report the sample sizes in some early studies was a major lapse in experimental rigour. However the final study discussed in detail later in this thesis employs two of the scenarios with similar results and reports appropriate sample sizes.

sure there is nobody else around. After deliberation, Fred Jerkson, because of his beliefs and values, keeps the money. After deliberation, Barney Kinderson, because of his beliefs and values, returns the wallet to its owner.

Given that, in this world, one's genes and environment completely cause one's beliefs and values, it is true that if Fred had been adopted by the Kindersons, he would have had the beliefs and values that would have caused him to return the wallet; and if Barney had been adopted by the Jerksens, he would have had the beliefs and values that would have caused him to keep the wallet. (2006:38-39)

Again, participants were required to answer a manipulation check, and were excluded from analysis if they failed to answer it correctly. Again, the results were similar to those of the previous experiments (2006:39). 76% of participants judged that both Fred and Barney acted of their own free will. 60% judged that Fred is morally blameworthy for keeping the wallet, and 64% judged that Barney is morally praiseworthy for returning it (sample sizes not reported).

The main conclusion that NMNT draw from these experiments is that most people do not take determinism to be incompatible with an agent's acting of his own free will or being morally responsible for his actions (2006:39). The common claim of incompatibilists is false, ordinary people's pre-theoretical intuitions do not support incompatibilism. NMNT do not go on to claim that the folk do have compatibilist intuitions. In their earlier paper they note that they consistently found a non-negligible minority of participants offering incompatibilist judgments (2005:572). They only briefly discuss this aspect of their results which, along with its counterpart in the study I discuss in section two, will be discussed in section four. They think that it may suggest that individuals have conflicting intuitions about free will or moral responsibility (what I will call intra-personal variation), or that individuals have internally consistent but externally diverse intuitions (what I will call inter-personal variation). They note that the former idea exists in the work of Richard Double (1991, 1996) and in the suggestion by Graham and Horgan (1998) that these concepts are context-sensitive. The latter may be foreshadowed in the work of William James and Ludwig Wittgenstein where it is suggested that the longevity of

the free will debate is explained by the fact that the belligerents start off with different views rather than falling into them in the course of the fray.

NMNT are careful to note that their results are preliminary rather than conclusive, and require further research on the distribution of intuitions about free will (2005:572). They note a major limitation of their study is that the participants were all college students, and suggest replicating it with subjects of diverse educational, cultural and socio-economic backgrounds. They also endorse taking a more qualitative approach involving coding responses in open-ended interviews and experimentally studying behavioral responses in situations involving responsibility attributions.

They also found in initial studies that participants had trouble reasoning conditionally. In their explanations of their responses they implied that the scenario is impossible because Jeremy has free will, presumably in a libertarian sense (2005:566). In response NMNT included first in all their experiments a question about whether the scenario is possible (in the following example this is ‘question 1’), and then asked subjects to reason conditionally, as for example:

Regardless of how you answered question 1, imagine such a supercomputer actually did exist and actually could predict the future, including Jeremy’s robbing the bank (and assume Jeremy does not know about the prediction).
(2005:566)

In this case the majority of the participants answered that the scenario was impossible, offering a wide range of reasons like the difficulty of procuring enough information and the truth of quantum indeterminism. Thus NMNT admit that some of the responses were probably not based on an acceptance of determinism (2005:574). In line with this some people failed the manipulation check, and although they were excluded from analysis there were likely others who made it through by chance rather than by really understanding the scenario. In fact they recognize that some of these people may have been libertarians who found the scenario impossible *because* they believed Jeremy had libertarian free will. Thus the results of this study may have been slightly biased towards suggesting that people

had compatibilist intuitions, given that some may not have believed that the scenario was deterministic.

An important problem in this research, as I mentioned in chapter one, is making the descriptions of determinism accessible to the participants without begging the question. NMNT first ran a pilot study that showed that the term “determinism” itself is not understood by the folk in the ways that philosophers use it (2005:565). It is usually interpreted by them as meaning having a set fate, being unable to choose, simply not having free will, or even as implying resoluteness. This alone does not imply that the folk are incompatibilists because it does not suggest that they understand the relevant concept of determinism. Thus NMNT think that describing a deterministic scenario according to a philosophically relevant and neutral concept, and without using the term “determinism”, is the best way to evaluate the relevant folk intuitions.

NMNT note that descriptions of determinism used in the literature, such as van Inwagen’s (1983, pp.58-65), that the past and the laws of nature entail any true proposition, are most precise but would not be accessible to participants without significant training in the relevant technical concepts (2005:572). On the other hand explaining determinism without using technical notions is more likely to include confounding factors that would be difficult to separate in analyzing the results. They consider but reject the strategy of describing determinism in terms of lacking the ability to do otherwise, as the notion of ability relevant to free will is controversial in the debate, and their research shows that the relevant folk theory is complicated. They welcome the attempt to strengthen their descriptions of determinism for further research, but they advise against philosophers judging the results of such research in advance from their armchairs.

Given that the understanding of determinism is achieved, it is still possible that participants not recognize the implications. NMNT distinguish a motivational and a cognitive form of this problem (2005:574). The motivational problem is that participants may be so attached to the existence of libertarian free will that they are hindered from making the inference that people might not be free in this way. This is similar to Peter Strawson’s (1962) view that the reactive attitudes that ground free

will cognition are impossible to change or eliminate. This would bias the results towards people having compatibilist intuitions because they are motivated not to judge that people do not have free will, even though they think that it is incompatible with determinism. But NMNT think that if such people understand the determinism in the scenario they will respond by inferring that agents in it are not like them. They suggest that people readily understand that people who are brainwashed do not have free will, presumably because they are not like them. So they do not think this is a major additional problem.

The cognitive problem is that people may be unable to easily comprehend the connection between free will and determinism (Nahmias, et al., 2005:575). It could be that the connection is not obvious from a description of determinism alone, but requires some kind of inference to be made. I think that one way to spell out this worry is that perhaps people would not see from a description of determinism that it rules out our having alternative possibilities, or being the ultimate source of our actions, in the ways that incompatibilists require for responsibility. In that case we could not conclude from their reporting compatibilist intuitions that they would not report incompatibilist intuitions upon seeing this connection. NMNT think that helping participants to comprehend the connection between free will and determinism is tantamount to offering arguments one way or another. They think that such arguments would contain premises that are themselves controversial and subject to appeals to ordinary intuitions. They think that this defeats the point of the incompatibilist claim to ordinary intuitions, constituting the strategy of talking the folk into incompatibilism rather than showing that they already are intuitive incompatibilists. However they note that it is empirically testable to what extent and in what ways learning about the free will debate changes people's intuitions. I think that they are right to be concerned about this, as they are aware that one of the main motivations for experimental philosophy is avoiding theory contamination, and biased presentations of issues based upon it. But I think that there may be scope for exploring the possibility of carefully testing how people understand descriptions of determinism. We could, for example, try to test whether people see a connection between descriptions of determinism and alternative possibilities or sourcehood, in

the incompatibilist senses. This is an important question for further research, but not one that I will try to answer in this thesis.¹⁷

2. *Nichols and Knobe*

The next study I will discuss was run by Nichols and Knobe (2007) and is probably the most influential experimental study on the compatibility question to date. Their aim was to begin to explore the psychological processes that underlie the production of intuitions about free will and determinism in order to assess whether they are generally reliable or not.

Like NMNT, Nichols and Knobe note that philosophers have traditionally claimed that incompatibilism is intuitive to the folk (2007:666). They note that philosophers have also claimed that people believe that human decisions are not governed by deterministic laws. While they think that the latter seems to have been vindicated by experimental results (such as Nichols, 2004) the same is not true for the former. Woolfolk, Doris, and Darley (2006) found that people found agents who were given a compliance drug, so that they had no choice but to perform an immoral action, were nevertheless more responsible for doing so if they wanted to than if they did not. The two scenarios were described as follows (2006:287-288 and 292): “The initial section of each vignette was common to all four conditions. In this section two married couples, Susan and Bill and Elaine and Frank, are depicted on a Caribbean vacation, and subsequently on board an airliner returning home. It is revealed that Susan and Frank have been involved in a love affair and that Bill has discovered proof of the affair.” In the ‘high-identification’ case this was followed by:

The humiliation and betrayal were almost more than he could bear. These were the two people he trusted most in the world. During the three days of the vacation that remained, he wrestled with the issue. He thought of many

¹⁷ It is discussed further in Nahmias, et al., 2006:43-48, and Sommers, 2010:204-206.

ways of retaliating. Finally, he decided that there was only one way he could deal with it. Bill decided that he would kill Frank.

In the ‘low-identification’ case it was followed by:

During the three days of the vacation that remained, he wrestled with the issue. Finally, he decided that if Susan and Frank wanted to be together, he would not stand in their way. He would confront them with the evidence and assume that whatever happened would be for the best. He really cared for both of them and wanted to be a forgiving person. He felt somewhat at peace with himself.

In both cases the final section read:

On the return trip home their plane was hijacked by a gang of 8 kidnappers. They then seized two of the male passengers, Bill and Frank. The leader of the kidnappers injected Bill’s arm with a “compliance drug”—a designer drug similar to sodium pentothal, “truth serum.” This drug makes individuals unable to resist the demands of powerful authorities. Its effects are similar to the impact of expertly administered hypnosis; it results in total compliance. To test the effects of the drug, the leader of the kidnappers shouted at Bill to slap himself. To his amazement, Bill observed his own right hand administering an open-handed blow to his own left cheek, although he had no sense of having willed his hand to move. The leader then handed Bill a pistol with one bullet in it. Bill was ordered to shoot Frank in the head. But when Bill’s hand and arm moved again, placing the pistol at his friend’s temple, Bill had no feeling that he had moved his arm to point the gun; it felt as though the gun had moved itself into position. Bill thought he noticed his finger moving on the trigger, but could not feel any sensations of movement. While he was observing these events, feeling like a puppet, passively observing his body moving in space, his hand closed on the pistol, discharging it and blowing Frank’s brains out.

In the low-identification case the mean score for Bill's responsibility was about two on a seven point scale, and in the high-identification case it was about three (12 participants were randomly assigned to each case). Nichols and Knobe think that this result fits well with the compatibilist view proposed by Frankfurt (1988), that an agent may be responsible for an action, because they identify with it, despite not being able to do otherwise. Identification, here, is concerned with certain types of desires. Frankfurt distinguishes first-order desires from higher-order desires. An example of a first order is the desire for an ice-cream. An example of a second-order desire might be the desire that one not have the desire for an ice-cream. Higher-order desires are defined in the same way. Identification is roughly the notion that an agent has consulted their desires, as well as their higher-order desires, and made a decision with which they are satisfied. Nichols and Knobe think that, as the responsibility ratings were still low in both conditions, this result does not pose a direct threat to incompatibilism, but that NMNT's (2005) experiment described in section one does (2007:667).

They think that it is possible to criticize that study on several grounds. The NMNT scenarios involve vocabulary such as "laws of nature" and "current state" which is technical and thus makes it unclear whether the participants understood them. I think that this is exacerbated by the fact that NMNT do not report what their manipulation check was or how many people failed it. Nichols and Knobe note that the scenarios also focus on the predictability of events in the universe, which many philosophers have thought to be less threatening to responsibility than causal inevitability. On the other hand they think that it is a positive quality of that research that it employed questions about moral responsibility rather than about free will alone.¹⁸ They think that the latter has become a term of philosophical art and thus devolved from the common interpretation of the folk, but that moral responsibility in the ordinary sense is often claimed to be the target of the free will debate. Nevertheless they think that NMNT are right that people tend to hold an agent morally responsible for intentionally performing an immoral action in a deterministic

¹⁸ There are also concerns about ambiguity in the concept of moral responsibility, which will be discussed in chapter three.

setting. They also note that independent psychological research has found that subjects identified as determinists are no less punitive and retributivist than others (for example Viney, et al., 1982), though they think that it is flawed in using a too liberal measure to identify determinists. However they suggest that people may have intuitions that are incompatibilist in some contexts but not in others.

They think that the folk find incompatibilism intuitive in the context of philosophical discussion and that this is why philosophers have so often claimed that the folk find incompatibilism intuitive (2007:668). They wondered whether questions that call for abstract theoretical cognition might elicit incompatibilist intuitions, while questions that trigger emotions might do the opposite. They note that it is somewhat counter-intuitive to think that philosophical examples would trigger strong emotions because they are generally thinly described hypothetical cases. But blossoming experimental evidence suggests the opposite. Some has also shown that when a transgressor is only slightly more clearly identified people feel more negative emotion towards them and are more inclined to punish them (Smart and Loewenstein, forthcoming). This leads Nichols and Knobe to note that all of the previous studies of folk intuitions about responsibility involved scenarios that featured determinate agents and thus lend themselves to causing negative emotions in the subjects. They hypothesized that incompatibilist intuitions are held as a folk theory but that compatibilist intuitions are produced by other psychological processes.¹⁹ They think that emotional responses might promote compatibilist intuitions by hindering people's attempts to apply their theories. They do not claim that such a folk theory would obviously be superior in terms of reliability but that further research would be required to find out. Later they note that one might think that emotional responses are required for competent judgments of moral responsibility, and present arguments against this view.

They set about exploring these possibilities by designing a series of experiments in which emotional factors are controlled and manipulated. Like NMNT they note that the most precise way to define determinism is by using technical language that is inappropriate to the study of folk intuitions, and that the translation to more simple

¹⁹ By theory they just mean some internally represented body of information.

terms is problematic (2007:668). However they note that the primary goal of their study was not to measure the commonality of incompatibilist intuitions among the folk but to discover whether emotional reactions influence their production. Thus it is not a big problem if their description of determinism tends to slightly bias the responses one way or the other. It would still be possible to see whether there is a significant difference between responses in abstract versus concrete and emotionally provocative contexts. While they do draw conclusions later about the existence of incompatibilist intuitions, as will come out shortly, they do not go so far as to say that the folk are natural incompatibilists.

Their experiments were based on the following description of deterministic and indeterministic universes:

Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries.

Now imagine a universe (Universe B) in which *almost* everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it *did not have to happen* that Mary would decide to have French Fries. She could have decided to have something different.

The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are

not completely caused by the past, and each human decision *does not have to happen* the way that it does. (2007:669)²⁰

They presented the descriptions to undergraduates at the University of Utah. They note like NMNT that it is important to check whether the results generalize to other populations but again their aim was primarily to test whether context has an effect.²¹ The description was followed by the question:

Which of these universes do you think is most like ours? (circle one)

Universe A

Universe B

Please briefly explain your answer: (2007:669)

They found that over 90% of participants responded that the indeterministic universe is more like ours (sample size not reported) (2007:669).

Distinct and randomly selected groups were then presented with one of the following questions that are abstract and concrete respectively:

In Universe A, is it possible for a person to be fully morally responsible for their actions?

YES NO

²⁰ There has been significant criticism of the use of the phrase “has to happen” in this description. This will be discussed in some detail in chapter three.

²¹ The abstract case in this experiment has since been employed in a cross-cultural study of students from the USA, India, Hong Kong, and Colombia, and found similar results. (Sarkissian, et al., 2010)

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.

Is Bill fully morally responsible for killing his wife and children?

YES NO (2007:670)

In the abstract case they found that 86% of participants said that an agent could not be responsible for their actions (sample size not reported). In the concrete case 72% said that Bill was responsible for his action (sample size not reported). They were concerned that this difference might be explained by the concrete case being so long and drawn out that the subjects lost track of the fact that the agent was in a determinist universe. To test this they presented another group with the following question:

In Universe A, Bill stabs his wife and children to death so that he can be with his secretary. Is it possible that Bill is fully morally responsible for killing his family?

YES NO (2007:670)

Here they found that 50% of subjects responded as compatibilists, which is a statistically significant difference from the abstract case (sample size not reported).

They claim that the high rate of incompatibilist responses to the abstract case cannot plausibly be attributed to a subtle bias in the description of determinism involved, because in the concrete case a significantly higher proportion gave compatibilist responses to exactly the same description. They take these results to replicate those of NMNT, but also to show for the first time that when presented with an abstract vignette subjects have predominantly incompatibilist intuitions (2007:671). Thus

affect seems to strongly influence the production of intuitions about moral responsibility.

In using the idea that given the past, each decision *has to happen* the way that it does, Nichols and Knobe aimed to test intuitions about the kind of compatibilism that they take to be most popular in contemporary philosophy (2007:682). That is the kind that follows Frankfurt in rejecting the requirement of alternate possibilities for moral responsibility. They note that some compatibilists instead reject the claim that determinism implies that people have to do what they do, those who do require the ability to do otherwise for moral responsibility, but only a conditional ability. They imply that this could not be too much of a problem given that in some cases most participants responded as compatibilists. I am not convinced that this may not introduce some bias into their results, but I will discuss this further in chapter three.

Nichols and Knobe set out several models that might explain the pattern of folk intuitions found. One they call the ‘performance error model’, that ordinary judgments regarding responsibility rely on a tacit theory but flagrant moral transgressions prompt strong feelings that undermine people’s ability to apply that theory (2007:671). This distinguishes people’s internal representations of the criteria for moral responsibility from the process by which they apply them, and asserts that emotions can interfere with the latter. It is supported by a large body of literature in social psychology on the interaction between emotion and theoretical cognition. This has shown that emotional and motivational biases can cause failure to recall relevant information, believe unwanted evidence, or use critical resources to evaluate conclusions (see Kunda, 1990, for a review). It is plausible that these biases apply to responsibility judgments also. Perhaps participants in these studies do not recall, or believe, the deterministic nature of the scenarios in judging agents responsible in emotional cases, or perhaps they do not carefully reflect on its implications.

There is also direct evidence that emotions affect responsibility judgments. Lerner et al. (1998) tested this by showing a video of a bully beating up a teenager to some people but not others. They then told them all that they would be doing another unrelated experiment, and asked them to rate the responsibility of a different agent who had demonstrated negligence in an unrelated scenario. Those people who had

watched the video of the bully gave higher responsibility ratings than those who did not. Lerner et al. conclude that the negative emotions aroused by watching the video of the immoral action caused people to make stronger judgments about the responsibility of an agent. As these emotions do not seem to be relevant to how responsible the agent is, they seem to be biasing people's judgments about it. Nichols and Knobe think that this kind of data supports the interpretation of their results in which there is no real sense in which people have compatibilist views of responsibility, but that concrete scenarios can cause performance errors resulting in seemingly compatibilist responses (2007:672).

Another possibility is that emotional responses, rather than abstract theories, actually constitute the most important part of how people ordinarily judge responsibility. Nichols and Knobe call this the 'affective competence model' (2007:672). It gains some support from evidence that people with psychological illnesses that inhibit their emotional processing sometimes exhibit bizarre patterns of moral judgments (see for example Blair, et al., 1997). This suggests to some philosophers that affect plays an important role in people's moral judgments (Haidt, 2001, Nichols, 2004a, Prinz, 2007). So it might be that people have a theory of responsibility that is in conflict with their actual judgments. It may be that their judgments spring not from their beliefs about under what conditions a person is morally responsible, which may be incompatibilist, but from emotional responses that bypass those beliefs, and which may be compatibilist.

The final possibility that Nichols and Knobe consider is that their results are explained by the concreteness in the scenarios rather than the emotional salience. They call this the 'concrete competence model' (2007:673). One appealing version of this view is that people have an innate moral responsibility module. Some people think that many basic cognitive functions, and moral functions in particular, are subserved by cognitive modules. These are mechanisms that are fast, automatic and inaccessible to consciousness. If a module took information about an agent and their behavior, but not conscious beliefs about the conditions required for responsibility, as inputs then it could deliver judgments that are quite out of line with those beliefs. People may believe that to be responsible an agent must not be subject to determinism. But their judgments about the responsibility of an agent may be

produced by a module that does not take determinism into account. In that case they may judge that an agent in a deterministic scenario is responsible for their actions, even though that is inconsistent with their other beliefs. Nichols and Knobe note that combinations of components of the models they discuss are also possible.

To begin testing which if any of these models could be true they ran an experiment to check whether emotion plays any role above and beyond concreteness (2007:675). They did so by employing two scenarios which were both concrete, but of which one was high affect and one low affect. Using the same description of deterministic and indeterministic universes they randomly assigned half of the subjects one of the following questions:

As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?

As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?
(2007:675)

They also stipulated for half of each group that the agent was in the deterministic universe and for the other half that the agent was in the indeterministic universe. Their results are presented in the following table in terms of judgments that the agent was responsible (sample sizes not reported) (2007:676):

	Agent in indeterminist universe	Agent in determinist universe
High affect case	95%	64%
Low affect case	89%	23%

Large majorities judged that the agent was responsible in the two indeterministic cases, a small majority in the high-affect deterministic case, and a minority in the low-affect deterministic case. There is a statistically significant difference between the high and low affect cases for the deterministic universe, even though both are

concrete, and the low affect case elicited mostly incompatibilist responses (2007:683). The difference between the deterministic and indeterministic cases was also statistically significant for both the high and low affect cases. Thus even when concreteness is controlled for emotion appears to play a significant role in judgments about compatibilism.

Nichols and Knobe think that these results support either an affective performance error model or an affective competence model (2007:676). They think that the performance error model is a more plausible explanation because it could be that in the low-affect cases the difference between the response rates is due to an incompatibilist theory held by the participants. It could be that in the high-affect cases this theory is being overcome by an emotional response in many of the participants, which causes them to judge that the agent is responsible despite their believing that determinism rules this out. They think that an affective competence account cannot explain why there is a big difference between the responses to the deterministic and indeterministic conditions in the low-affect cases, but not in the high-affect cases. They think that explaining how our competence in attributing responsibility is undermined in the low-affect cases would “take significant work” given that cheating on one’s taxes is relatively ordinary. They conclude that these results *prima facie* favor the performance error account, but that deciding the issue will require a deeper understanding of the role of emotion in moral cognition.

One of the most exciting inferences that Nichols and Knobe make from these results is the possibility of explaining the persistence of the free will debate (2007:677-678). It seems that different psychological processes favour different answers to the compatibility question. Furthermore, if one of those processes is unreliable then we would have good reason to ignore or discount the intuitions associated with it. If the performance error model is right then if the errors are eliminated it might be that the compatibilist intuitions disappear. This would imply an error-theory for compatibilist intuitions, showing that they would not be evidence for compatibilism. If the affective competence model is correct then the normal system for assessing responsibility is a compatibilist one. This is similar to P.F. Strawson’s view that the concept of responsibility is not centrally a metaphysical one but a system of reactive attitudes that underlie our responsibility practices. If the concrete competence model

is correct and there is a responsibility module that takes concrete information as its only input it could be that other important theoretical information is ignored, leading to a kind of moral illusion. On the other hand the module itself could contain complex and useful information that is not present in folk theory.

Nichols and Knobe ran another experiment to see which intuition people would prioritise, the abstract incompatibilist one or the high-affect compatibilist one. They presented subjects with the same description of universes but asked them both of the relevant questions (abstract and high-affect questions labelled as questions 2 and 3 respectively), counter-balanced for order. The subjects then read:

Previous research indicates that when people are given question 3 above, they often say that Bill is fully morally responsible for killing his family. But when people are given question 2 above, most people say that it is not possible that people in Universe A are fully morally responsible for their actions. Clearly these claims are not consistent. Because if it is not possible to be fully morally responsible in Universe A, then Bill can't be fully morally responsible.

We are interested in how people will resolve this inconsistency. So, regardless of how you answered questions 2 and 3, please indicate which of the following you agree with most:

- i. In Universe A, it is *not* possible for people to be morally responsible for their actions.
- ii. Bill, who is in universe A, *is* fully morally responsible for killing his family. (2007:683)

They found that roughly half (10/19) agreed most with the incompatibilist judgment, and half with the compatibilist judgment. They do not take this as evidence about which is more reliable, but they think that it is interesting that making a simple attempt to seek a personal reflective equilibrium does not appear to resolve this issue. If it did it might mean that one of the intuitions would govern how people

would actually make responsibility judgments upon considering both abstract and concrete cases.

Nichols and Knobe conclude that while philosophers who discuss folk intuitions tend to say that they are either compatibilist or incompatibilist, it appears that people have both intuitions, and that they are produced by different psychological mechanisms (2007:681). How important this is to the debate depends on understanding those mechanisms and their reliability in detail, which they only claim to have started in their work so far. I am broadly in agreement with Nichols and Knobe's assessment of their own study, and of NMNT's earlier studies, except where I have stated otherwise. In the next section I will present some discussion of how to interpret these studies, and in section four I will make some contributions to that discussion.

3. *Doris, Knobe, and Woolfolk*

Doris, Knobe, and Woolfolk (henceforth DKW) (2007) think that the studies described above suggest a problem for the standard approach to the free will debate. They note that the standard methodology relies a good deal on folk intuitions (2007:183). They also note that progress towards agreement in the debate is uncertain. They think that the experimental data being collected shows that these facts are related. Intuitions will not settle the debate because they support different theories in different cases.

DKW explain this view in terms of two assumptions that are common in the free will debate but that they think cannot both be true. They call them 'invariantism' and 'conservativism'. Invariantism is the view that there are exceptionlessly relevant criteria for an agent's being judged responsible for what they do (2007:184). Incompatibilists often require that an agent have alternate possibilities or be the ultimate source of their actions. Compatibilists often require that an agent identify with their action or intention, be able to respond to right reasons to act, or be free of ignorance or coercion. Each of these requirements is supposed to apply in any context, so they are all invariantist. This is consistent with their being complex, vague, or based on prototypes or paradigms rather than necessary conditions. But invariantism is more than applying rules or principles; it is applying the same criteria

to all cases. There may be theories which fall somewhere in between applying the same criteria to all cases and applying different criteria to different cases. But the major theories in the free will debate clearly assume invariantism, though it is not often explicitly stated.

Conservatism is the view that folk belief is a (defeasible) constraint on philosophical theorizing (2007:185). In the free will debate this tends to include folk morality in general, including beliefs, judgments, practices, and (perhaps inchoate) theories. This assumption is more often explicitly stated, as shown in chapter one. DKW note that one view in the free will debate clearly rejects this assumption. Hard determinism, as described earlier, is the view that determinism is true and rules out our having free will and moral responsibility. As free will and moral responsibility are very important to us, and underlie a great deal of folk morality, hard determinists seem to reject conservatism. However hard determinism has always been relatively unpopular, which is probably because it seems to reject conservatism while most philosophers endorse it.

DKW think that invariantism and conservatism may be individually plausible, but that they cannot both be satisfied (2007:185). As folk morality is variantist, philosophical theories that are invariantist cannot be conservative. DKW think that the data collected in the studies described above, along with other data on the attribution of responsibility, raises this problem by strongly suggesting that folk intuitions regarding responsibility indicate the application of different criteria in different cases. While DKW avoid speaking in terms of free will, the sense of free will that I am discussing in this thesis is that of the conditions required for moral responsibility, as I have mentioned. This is not to be confused with a claim about the analysis of the folk concept of responsibility. DKW think that such claims rely on some account of concept structure and individuation, and that such accounts are subject to controversy in both philosophy and psychology (2007:186). The concept of responsibility may be simple and straightforward, while the conditions of its application vary across different cases. Later, as I will mention, they consider the ways that terms to do with responsibility might be interpreted by the folk, which some philosophers might call concepts. But they avoid talking about concepts specifically. They think that the data collected may not provide conclusive evidence

for claims about the nature of the concept of responsibility. But they think that the studies described above suggest that people apply compatibilist criteria for responsibility in some cases and incompatibilist criteria in others. They also think that other studies have suggested that, contrary to philosophical theories, people require different mental states in an agent in different cases, for them to be responsible. I will be focusing on their views regarding folk intuitions about compatibilism and the implications of their being variantist. I will present them in this section and set out my own view in the next.

DKW note that many compatibilists have been inspired by the work of Peter Strawson (1962) in claiming that the problem of free will and determinism is an invention of philosophers that is not supported by folk morality (2007:187). Incompatibilists reject this view, claiming that folk morality requires that determinism be false for agents to be responsible. Thus DKW support the picture of the role of folk intuitions described in chapter one. Both sides of the free will debate endorse the view that theories that go against folk morality thereby take on a burden of proof, and that their own view is supported by it. DKW also agree that the way to resolve this conflict is by conducting experimental studies.

They suggest a slightly different analysis of the two experimental studies reported above (2007:191). They think that NMNT's bank-robbery case is relatively less affectively laden than the high-affect cases in Nichols and Knobe's study. However, as reported in section one, the majority of participants gave apparently compatibilist responses to the bank-robbery case. DKW think that this may count against the restriction to affect-based models endorsed by Nichols and Knobe. What these cases seem to have in common is that they all ask about a particular agent who has performed a particular action. They think that this suggests that a concreteness-based model better explains the results. People may be inclined to apply compatibilist criteria for responsibility in cases involving a particular agent, but incompatibilist criteria when considering agents in general.

In another paper Knobe and Doris note that the usual view in philosophy is that it can be discovered whether a principle fits with ordinary judgments of responsibility by asking people whether they agree with the principle itself, and by asking them

about particular cases that are relevant to the principle (2010:326). This is an important step in reaching a reflective equilibrium. But cognitive science shows that these often actually lead to quite different results. The principles people assert in abstract conversations can have nothing to do with how they actually judge concrete cases. The point here seems to be that asking someone whether people could be responsible for their actions in a deterministic world, and asking them whether a particular person in a particular situation is responsible for their actions, may induce responses that are inconsistent with each other.

DKW think that a difference in perspective between compatibilists and incompatibilists may show that individual philosophers have simply been focusing on one or other of these contexts (2007:191). Compatibilists tend to follow Peter Strawson's lead in considering everyday moral practices. Incompatibilists on the other hand, focus on the intuitions that are more common in the abstract context of philosophical discussion.

They take this to make it tempting to think that folk morality is incoherent in these cases, because the concreteness of a scenario seems like a factor that is irrelevant to the attribution of responsibility. They note that there is evidence that moral intuitions can be influenced by morally irrelevant factors such as how a question is framed (for example Doris and Stich, 2005). It is thus plausible that there is some inconsistency or error involved in some of the results, and that such mistaken judgments should be removed from philosophical analyses. This would suggest that there is less variation than at first appears, as some of that variation would just be mistakes, and not what the participants really think. But DKW think that from a certain perspective it actually seems reasonable to invoke different criteria for moral responsibility in different contexts. In philosophical contexts it seems appropriate to worry about abstract metaphysical theses such as determinism. But in everyday personal interactions that is unnecessary and inappropriate. So there may be some general intuitive backing to the conflicting pattern of intuitions found in these studies.

DKW think that there are two general ways of denying the empirical claim that folk intuitions about responsibility are variantist (2007:197). One is to deny that the data

cited is reliable evidence of the nature of folk morality. One way to do this is to deny that experimental philosophy can tell us people's real moral judgments. DKW note that the validity of experimental philosophy has been defended elsewhere (for example in Doris and Stich, 2005) and do not seek to add to this debate. Instead they note that thought experiments are widely used in philosophy and that experimental philosophy simply applies this approach with greater scientific rigor.

Another way to attack the data is to deny that it is evidence of variation in responsibility judgments. One concern, DKW think, is that some of the studies ask about praise and blame rather than responsibility directly (2007:198). Perhaps these judgments are more likely to be subject to variation of the relevant kind. But firstly there is empirical evidence that they are highly correlated with responsibility (for example in Alicke, 2000). And secondly, as in Strawson's work linking responsibility attribution to reactive attitudes, philosophers frequently take praise, blame, and responsibility to be closely related.

This touches on an important point. It may be that people have several concepts of responsibility, and that there is variation in when people employ one or another of these concepts. For example it could be that people sometimes use a concept of responsibility as being causally involved in the occurrence of an event, and sometimes as being an apt target for reactive attitudes such as resentment or gratitude, praise or blame. In that case the attribution criteria for each of these concepts may be invariant, but they may be different for each concept. The variation in the results of these studies may then be due to people employing different concepts in different contexts, rather than the attribution criteria for one of those concepts being variantist. This would be interesting in itself, but if there were just one concept of responsibility that philosophers are concerned with, then this would not show that that concept is not attributed invariantly.

DKW seem to address this concern when they note that one may deny the efficacy of the experiments more generally by claiming that the variation exhibited is merely verbal, caused by different interpretations of subtle terms involved. DKW think this is a reasonable concern but that experimenters are motivated to avoid such problems (2007:198-199). To show that it does explain the apparent variation would itself

require careful experimentation. Again, they think that this is just as much a problem for traditional philosophical discussion, so it is not a reason to reject the experimental method. There is more empirical research to do, but as the evidence stands now, they think that variantism is the most likely state of folk intuitions about responsibility.²²

As to the second kind of objection to the data, it might be argued that while some aspects of the criteria applied in responsibility judgments vary, in all cases they involve considerations of causation or intentionality (2007:199-200). Thus while superficial variation may exist, this objection may run, these deeper criteria are exceptionlessly relevant. DKW think that this may be true, but that there is evidence that the criteria for attributing causation and intentionality are themselves variant. They relate the criterion of intentionality to the issue of the mental states required for an agent to be judged responsible. Concerning causation they cite studies (for example Alicke, 2000) that show that whether or not someone's behavior is judged to be the cause of something often depends on its moral value. And again they note that philosophers have suggested this view (for example Feinberg, 1970:200-207). DKW give the example that it may seem natural to say that Hitler's invading Poland was more properly the cause of World War II than Chamberlain's pledge to defend his allies, and this may only be because Hitler's action was wrong or blameworthy. They note that there may be other plausible ways to show that the data does not represent truly variantist folk intuitions. But given the evidence it will be difficult to do so, and a natural interpretation is that the variation observed is due to real variation in folk morality.

Thus DKW think that either conservatism or invariantism must be rejected for responsibility. This leaves two views. One is to reject conservatism, leaving invariantist theories that propose revisions to folk intuitions which vary in different contexts. DKW call this 'revisionary invariantism'. Otherwise invariantism must be rejected, leaving variantist theories that conserve folk intuitions that vary in different contexts. DKW call this 'conservative variantism' (2007:202).

²² I discuss this possibility in more detail in chapter four.

The aim of this thesis is to show that the experimental data discussed raises some problems for standard assumptions in the free will debate. I will briefly present DKW's discussion of how to deal with this problem in order to further illuminate the kind of problem it is. They think that the cost of revisionary invariantism will vary in different areas. In technical areas like mereology they think that the folk may be happy to let the experts overrule their own weakly held intuitions on the matter. Regarding responsibility they think that revision may at first appear most attractive, but will actually be difficult. They favor conservatism aimed at preserving widely-shared common sense (2007:203). They allow that reflection on the biological (for example Stanovich, 2004) and cultural (for example Railton, 2003) origins of folk morality suggests plausible genetic arguments in favor of revising some aspects of folk morality. Revisionism can come in different degrees, from eliminativism advocating rejecting the concept of responsibility altogether, to moderate revisionism (for example Vargas, 2005) which only modifies or removes relatively few concepts while retaining the bulk of folk theory. DKW think that the more moderate the revisionism, the more attractive it will be. But given the variation in the data, and the fact that normative considerations probably underlie much of it and are probably themselves highly variant, folk morality is probably pervasively variantist. While some aspects of the variation may be reflectively unified, any invariantist account will be extremely revisionist.

It is possible that extreme revisions could be justified so DKW consider the constraints on revision and the motivations for it (2007:204). Vargas (2004:230-233) suggests two constraints on revision. One is normative adequacy, which is that the revision must be "justified and integrated with our network of mutually supporting norms and practices". The other is naturalistic plausibility, which is that the revision must be plausible "under some broad-minded conception of substantive naturalism." The second has little to do with variantism. Whether the same criteria are applied to all cases, or different criteria in different cases, they may equally well be naturalistically plausible. As for normative adequacy, it could be argued that folk morality includes a strong commitment to invariantism. DKW suggest that something like "treat like cases alike" may be a plausible doctrine. But this is not violated by variantism, which treats different cases differently; invariantism itself is not a commitment of folk morality.

If enough individual variations were shown to be normatively inadequate then variantism in general may be more likely to be so. But DKW think that the ones they discuss are neither clearly normatively justified, nor clearly unjustified. Again, they are happy to allow that some may not be, but there is considerable work required to show that most are. Given that folk morality is pervasively variant it is invariantism that is subject to, and at risk of failing to meet the constraint of normative adequacy.

Even if revisionary invariantism is relatively poor in terms of normative adequacy it may be relatively well motivated. If philosophical or scientific theory better resolves the issues that folk theory deals with then that theory may enjoy what DKW call “revisionary prestige” (2007:205). One way that it may do this is by being instrumentally successful, which is often explained in terms of predictive and manipulative success. For example scientific medicine is better at treating illnesses than the former folk medicine, and thus is an obviously justified revision to it. Philosophical theories are less obviously instrumentally successful, in fact it is hard to say what this would mean. One way that invariantism might be relatively instrumentally successful is by being easier to use. In applying the same criteria to all cases it is a relatively simple approach, but in doing so it risks conflicting with strong commitments about particular cases.

Another source of revisionary prestige is explanatory success. Folk theories may be revised if they cannot explain things that others can. But revisionary invariantism seems to have a greater burden in explaining away the observed variation in folk intuitions about responsibility. Individual variations could be shown to be caused by errors or biases in judgment. But DKW think that there is no uncontroversial standard of what such an error would involve. They note that to know what such an error would be like we would need to have a reasonable understanding of the cognitive processes underlying these judgments. While this is a burgeoning area of research we plausibly do not yet have this understanding. Again while there may be errors that explain some of the variation discovered, they are not obvious.

Finally revisionary prestige may be gained by robust expert consensus (2007:206). But in the responsibility literature there is no such consensus. Rather discussions seem to be at an impasse. While there does seem to be agreement that invariantism is correct for responsibility, this is less of a reasoned consensus, and more of an unquestioned assumption. This is no reason to revise folk morality, but a reason to carefully consider the justification of the assumption. Given the underlying intractable disagreement involved such an assumption on first glance has little to recommend it.

DKW think that a positive case for conservative variantism may be difficult to make (2007:206-207). For one natural way to rationally justify a set of judgments is to find general principles that unify and ground them. But this goes against the core of variantism. DKW suspect, on the incomplete evidence at hand, that while there will be general principles that account for much of folk morality, they will vary greatly in their applicability. This is not just the point that any account of folk morality would be highly disjunctive. For invariantism is consistent with applying a highly disjunctive set of principles to judgments. What invariantism would require is that such a disjunction apply in the same way in any different context, and this is what DKW doubt to be true of folk morality. Whether this state makes good sense, or is normatively defensible, is a question for ethical theorists, and one that can only be answered on the basis of a detailed understanding of it. DKW conclude that conservative variantism deserves serious consideration.

4. Variation in Folk Intuition

Contrary to DKW, I think that Nichols and Knobe's experimental study suggests that emotion is influential in many folk judgments about the compatibility question, though it is not clear whether this is in terms of producing an error. The fact that significantly fewer participants attributed responsibility in the case of someone cheating on their taxes than in the case of someone stalking and raping a stranger seems plausibly to be due to the latter arousing much stronger emotions. DKW's suggestion that a concreteness-based model is vindicated by NMNT's bank-robbery case is unpersuasive. Many people would have a strong emotional reaction to the idea of someone robbing a bank. While it is plausible that it would cause less

emotion than Nichols and Knobe's case of stalking and rape, the fact that the majority of participants attributed responsibility in it could be due to the different descriptions of determinism involved in the studies. In the next chapter I will argue that Nichols and Knobe's description may cause some compatibilist folk to respond as incompatibilists. This may mitigate the effect of the relatively strong emotional salience of the scenario. There is also the possibility, described in section two, that the Laplacean description of determinism employed in that study did not make determinism sufficiently salient to the participants. Another explanation could be that emotion does not have a linear relationship with responsibility judgments. Rather than it being the case that the more emotionally salient a scenario the greater the proportion of responsibility attributions it will provoke, it may be that once a certain level of emotional salience is reached everyone who will attribute responsibility in response to it does so. In that case the emotion aroused by the bank-robbery case may be sufficient to reach this limit and though the rape scenario provokes stronger emotions they may have no further effect on people's responses.

A better example with which to make the argument that emotion cannot account for the responses to all of the different scenarios may be the scenario involving the neutral action of going jogging in NMNT's earliest study.²³ This plausibly would not have aroused any significant emotions in the participants at all, and yet it received similar responses to the other concrete cases (79%, 68%, and 76% of participants judging that Jeremy goes jogging, saves a child, and robs a bank, of his own free will, respectively). However this does not settle the matter in favor of a concreteness model. Nichols and Knobe's experiment in which they control for concreteness and manipulate emotional salience suggests that emotion can influence judgments even between concrete cases. The concreteness model cannot account for the low level of responsibility attributions to the low-affect action of cheating on taxes. This is an interesting puzzle and I do not think that there is an obvious solution on the basis of the results of these studies. Further research is required. Other arguments in favor of a concreteness-based model will be considered in the next chapter.

²³ Nahmias and Murray (2010) raise this point, as I discuss in chapter 3.

I think that DKW are right that if concreteness does effect folk intuitions about responsibility there is no good reason to think that there is an error on either side. It is clearly incoherent to hold both that no one could be responsible in a deterministic world and that a particular person is responsible in that world. But it is also possible to unknowingly hold incoherent beliefs, or have incoherent intuitions, which is why it is one goal of philosophers to identify and revise away such conflicts. I will not try to evaluate DKW's arguments that revision may not be the best option in this case. I am interested in the questions raised by these experimental results, but I think that they will take a lot of research to answer. But I think that their arguments can be made to work just as well, or even better, for an emotion-based explanation of the results. Their argument that abstract versus everyday contexts may appropriately elicit different criteria for responsibility judgments, is based on concreteness accounting for these conflicts. But ordinary personal interactions are relatively emotionally salient as well as relatively concrete, compared with philosophical discussions. It is a common and plausible view that it is good to be able to step out of one's emotional state and judge the moral aspects of situations objectively. This is exactly the kind of thinking one might expect to be elicited in abstract philosophical contexts. On the other hand it is also relatively common, at least among philosophers these days, as Nichols and Knobe point out, to think that feeling emotions is sometimes required in order to understand moral situations. The apparent relation between the lack of certain emotional responses and normal moral judgments in psychopaths seems to support this view. Perhaps we must be able to step into other people's shoes, not to be more objective but to imagine how they would feel and react, in order to really morally understand situations.

DKW's argument that we do not yet understand the nature of conceptual cognition well enough to really know what an error in it would consist of seems to apply equally well or better to the role of emotion within it. This point is similar to Nichols and Knobe's concession that we do not yet understand the role emotion plays well enough to say whether a performance error model or an emotional competence model is correct. I agree with both points, the research at this point is inconclusive. Thus, whether they are due to differences in the concreteness or emotional salience of scenarios, it is not obvious that the conflicts in intuitions discovered in the studies above are explained by some error on either side.

Most of the discussion of the studies described in sections one and two has focused on the apparent fact that most people have intuitions that support incompatibilism in some contexts and compatibilism in others. I will call this ‘intra-personal variation’, because the point seems to be that, among many people, the same person has varying and conflicting intuitions. One aspect of the experimental results that is not much discussed is that when people are presented with the same scenario some people have intuitions that support incompatibilism and others have intuitions that support compatibilism. I will call this ‘inter-personal variation’, because the interesting point is that, for all of the scenarios employed, different people have different and conflicting intuitions. In no case did 100% of the participants respond in favor of incompatibilism, and the same is true of compatibilism. NMNT briefly mention the existence of significant minorities in their cases, but most commentators seem to find this less interesting than the intra-personal variation that DKW discuss.

If there is inter-personal variation in intuitions about free will and determinism that would be contrary to, or at least in tension with, another assumption that philosophers in the free will debate often make. That is that the folk generally share one set of conditions that they believe are required for responsibility. This is shown in the references set out in chapter one where philosophers on both sides claim that most people intuitively agree with them. It is also contained in John Martin Fischer and Mark Ravizza’s account of their methodology where they “...suppose that there is enough agreement about these matters – at some level of reflection – to justify engaging in the attempt to bring out and systematize these shared views” (1998:11).

It is unclear what level of inter-personal variation would be in conflict with these claims. Surely no philosophers think that everyone must agree with a philosophical theory for it to be correct. Surely, also, philosophers do not have any particular proportion in mind that sharply distinguishes acceptable levels of agreement among folk intuitions and a philosophical theory. But, to take a strong example, in NMNT’s experiment in which Fred and Barney’s beliefs and values are described as determined by their genes and their upbringing, only 60% and 64% of participants judged that they were morally blameworthy and praiseworthy respectively. This case is arguably one in which the determinism is described most realistically,

accessibly, and in most detail, and where moderate but not insignificant levels of emotion might be aroused. It is certainly not clear from these results that ‘ordinary people are incompatibilists’, or that ‘compatibilism is the default position’.

Feltz and Cokely (2008) have argued for the existence of inter-personal variation in intuitions about free will. They presented 58 participants with the following scenario which they borrowed from Nahmias et al. (2007):

Most respected psychologists are convinced that eventually we will figure out exactly how all of our decisions and actions are entirely caused. For instance, they think that whenever we are trying to decide what to do, the decision we end up making is completely caused by the specific thoughts, desires, and plans occurring in our minds. The psychologists are also convinced that these thoughts, desires, and plans are completely caused by our current situation and the earlier events in our lives, and that these earlier events were also completely caused by even earlier events, eventually going all the way back to events that occurred before we were born.

So, once specific earlier events have occurred in a person’s life, these events will definitely cause specific later events to occur. For example, one day a person named John decides to kill his wife so that he can marry his lover, and he does it. Once the specific thoughts, desires, and plans occur in John’s mind, they will definitely cause his decision to kill his wife. (2008:3-4)

They then asked participants to indicate on a 7 point Likert scale (from strongly agree to strongly disagree) their agreement with the statements: “John’s decision to kill his wife was “up to him.””, “John decided to kill his wife of his own free will.”, and “John is morally responsible for killing his wife.” (2008:4).

They also tested each participant’s personality for the trait ‘extroversion’ using a brief version of a psychological test called the Big 5 personality inventory (Gosling, et al., 2003). An extrovert, in the Big 5 model, is a “communicable, sociable, energetic person who thrives on social contact and who does not regulate tightly his/her emotional reactions” (2008:4). Extroversion is largely stable across a

person's life, and is associated with certain cognitive and behavioural sensitivities (2008:3).

Feltz and Cokely found similar proportions of participants reported compatibilist intuitions as in other concrete scenarios discussed in this thesis. 71% of participants judged that John's decision was up to him, 69% judged that he decided of his own free will, and 76% judged that he is morally responsible for killing his wife (2008:5) (compared, for example, with 76% and 83% judging that Jeremy acts of his own free will and is morally blameworthy for robbing a bank). But they also found that there was a strong correlation between extroversion and compatibilist intuitions. The more extroverted a participant was the more likely they were to have compatibilist intuitions.

Feltz and Cokely imply, I think correctly, that the fact that people's tendencies to have compatibilist intuitions are related to stable personality differences gives us more reason to think that there is real inter-personal variation in intuitions about free will (2008:2-3). They also draw the conclusion that philosophers have been wrong in claiming that the folk all have the same view about free will, and that this may help to explain the intractability of the debate. They go on to argue that the relationship between extroversion and compatibilism gives us an important clue about the psychological processes that underlie intuitions about free will, both in the folk and in philosophers.

How philosophers should respond to inter-personal variation is a complicated question. I cannot answer it here but I can suggest some options to illustrate the kind of issues involved. One option is to relativise the debate to people according to their intuitions. Philosophers could reject the assumption that there is just one ordinary way to attribute responsibility to people that drives the debate. Ordinary thinking may be divided between those who attribute responsibility on the basis of criteria that are compatible or incompatible with determinism. In that case the free will debate could be applied differently to different people. Those who are compatibilists would not have to face the problems the free will debate primarily deals with. If determinism is true for human behaviour they could consistently continue to attribute

responsibility to people. Incompatibilists would still have to consider the possible inconsistencies in doing so, and how to resolve them, as sketched in chapter one.

One reason that philosophers think that intuitions about free will are important seems to be that they are supposed to be widely-shared. So another way that philosophers could respond to inter-personal variation in intuitions about free will is to argue that the intuitions that are most common are the most important. Notice that this seems to underlie DKW's case for conservative variantism. Since the majority of people appear to have different intuitions in different cases, to support them requires rejecting the assumption that either compatibilism or incompatibilism is true in all cases.

Another possibility is that both compatibilist and incompatibilist intuitions are common enough to be important. Rather than only focusing on the majority intuition, philosophers might find more or less important aspects of both majority and minority intuitions. Perhaps compatibilism and incompatibilism both have some practical value. On the other hand, another possibility is that neither compatibilist nor incompatibilist intuitions are common enough to support philosophical theories. So philosophers may take inter-personal variation to count against intuitions about free will telling us anything useful and important. This view has been supported by experimental philosophers in other areas, as I mentioned in chapter one.

Feltz and Cokely take a more sophisticated approach, in seeking to find factors that explain inter-personal variation in intuitions about free will. They want to study the psychological processes that underlie those intuitions. Perhaps understanding those processes will give us reasons to prioritise one intuition over another. But, again, it could turn out that both compatibilist and incompatibilist intuitions, or neither, have reasonable grounds. It might even be that as we find out more about the sources of different intuitions about free will we uncover more diverse variation, either intra-personally as DKW suggest, or inter-personally.

I think that issues to do with intra and inter-personal variation motivate further research on the variation in intuitions discovered in the studies above. As this research has only just begun there are many possible explanations in terms of errors

that would undermine the experimental data that raise these issues. Those that are plausible should be experimentally tested. In the following chapter I will consider one that has been.

Chapter 3: The Bypass Error Hypothesis

In this chapter I will present a very recent experimental study that tests an error-based explanation of variation in folk intuitions about free will. I will argue that it does not provide significant support for that explanation. In section one I will present an early proposal of that explanation. It consists of two claims. The first is that incompatibilist folk intuitions are based on the mistaken inference that an agent's mental states have no effect on their behaviour in deterministic scenarios. The second claim is that in concrete scenarios this inference tends to be cancelled out by consideration of an agent performing an intentional action. If this hypothesis were true it would explain away both inter-personal and intra-personal variation in folk intuitions about free will. Thus it would undermine the questions about assumptions in the free will debate raised in chapter two.

In section two I will describe the experimental study that is claimed to support this hypothesis. It used two descriptions of determinism from the studies reported in chapter two but includes questions designed to determine whether people think agents have no effect on or control over their behaviour, as well as questions about moral responsibility. It found a strong correlation between responses to these questions. Most people who reported incompatibilist intuitions also judged that agents had no effect on or control over their behaviour under determinism, and most people who reported compatibilist intuitions did not. This was taken to show that most incompatibilist intuitions are based on the error of thinking that causation implies compulsion.

In section three I will argue that there are problematic ambiguities in the questions about effects and control. They are ambiguous between simple effects or control, and effects or control that is not determined by prior conditions. I will argue that incompatibilists are likely to interpret the questions in terms of the latter, because it accords with what they think is required for an agent to be responsible for their actions. Thus their responses may only indicate that they realise that agents do not have effects on or control over their behaviour that is not determined by prior conditions in a deterministic world. This would not imply that they are making an error. This strategy accords with a common incompatibilist response to the

charge of confusing causation with compulsion. That is, that it is not the false belief that determinism would imply that our mental states would not affect our behaviour that concerns them, but the fact that under determinism all of our mental states would themselves be caused by conditions prior to our birth, which we have no control over. I argue that the questions in this study do not show that the former, rather than the latter, is the case. Further research is required to find out.

In section four I will briefly present another plausible invariantist explanation of the variation in folk intuitions found. It is based on an ‘ambiguity thesis’ that there are several concepts of responsibility that are available to the folk, and that have different application criteria. Some examples are moral, legal, and causal responsibility. This might explain the results in terms of different people, or different contexts, tending to evoke the use of different concepts of responsibility, and those concepts being variable in their applicability to different cases. For example, it could be that in abstract cases people tend to employ a causal concept of responsibility. In that case they might judge that an agent is not the undetermined cause of their action, under determinism, and thus not attribute responsibility to them. In concrete cases they may employ a moral concept of responsibility so that they judge that an agent is morally responsible despite being subject to determinism. I will claim that this kind of thesis should be tested for in further research.

1. Nelkin

Nelkin notes that the data collected since NMNT’s early research described above has been recalcitrant to both compatibilist and incompatibilist analyses (2007:244). So much so that it is tempting to think that our concept of responsibility is more complex than philosophers have thought, or even completely confused. But she thinks DKW are wrong to think that folk intuitions about free will are subject to intra-personal variation.

Nelkin agrees that compatibilism and incompatibilism are invariant views, in DKW’s sense, as they offer one set of criteria for responsibility regardless of speaker or subject contexts (2007:245-246). And she agrees that philosophers have typically assumed that whatever theory is correct will be invariantist. She notes that this is

consistent with there being degrees of responsibility, and borderline cases where it is not clear whether someone is responsible for an action or not. An invariantist theory can also have disjunctive components at some level, so long as it is generally unified. As an example she considers the ability to recognise moral reasons to act, which many philosophers think is required for moral responsibility. Such an ability may involve cognitive or perceptual abilities that may be manifested in different ways. For example recognising a moral fact may be possible in different ways. Thus a theory that requires the ability to recognise moral reasons to act would be unified by that requirement, while being disjunctive in allowing that it be manifested in different ways. This seems plausibly to be an invariantist view because the unity has some principled grounds. Nelkin also notes that it is possible to be a prescriptive invariantist, holding what DKW call revisionary invariantism, that our ordinary judgments are variantist but that we should revise our concepts to be invariantist. Manuel Vargas (2006) takes this view of the experimental results for compatibilism that DKW consider.

Nelkin thinks that the data found in the experimental studies discussed in chapter two does raise a problem, because it shows that there is no simple explanation of responsibility judgments from a unified set of criteria and obvious features of scenarios (2007:251). But she does not think that this conclusively shows that an invariantist theory cannot explain most responsibility judgments. Firstly because, as noted in chapter one, philosophers who rely on appeals to intuitions typically aim to find a reflective equilibrium between intuitions and general principles, acknowledging that there may be some conflicts in our intuitions. Thus the fact that some initial intuitions are difficult to accommodate in an invariantist theory does not rule out all such theories. It is possible that conflicts in intuitions can be resolved by weighing up which are most important or by showing that some are based on an error. Secondly judgments about responsibility depend on background empirical assumptions and features of scenarios as understood by the subjects. These may well be different from what the experimenters expect, and can be subject to mistakes. Nelkin takes these points to ground a general strategy for invariantists to respond to apparently variantist data. Her point seems to be the reasonable one that the studies are preliminary, and that the results admit of several interpretations that may be tested by further research.

Nelkin suggests a general hypothesis about how the assumption of invariantism may be preserved in the face of the experimental evidence about responsibility (2007:255). It consists of two claims. The first claim is that the participants who judge that agents could not be responsible in a deterministic scenario, but not other participants, make a mistake in interpreting determinism. When they consider determinism on its own, even when it is described rather than labelled, they take it to imply that an agent's mental states do not affect their behaviour. The second claim is that this inference tends to be cancelled out when a particular agent performing a particular intentional action is described.²⁴ If this hypothesis were true it would show that the intra-personal variation that DKW discuss does not exist. It is especially interesting because it would also rule out the inter-personal variation that I discuss, based on the minority responses in each kind of scenario. It can do this because it posits an error based on the description of determinism, rather than on the abstract, emotional, or concrete nature of the scenario, and which is thus plausibly present in all of these cases.

Nelkin suggests two ways that such a mistake could be made. One is the common idea among compatibilists that incompatibilists mistake determinism to imply coercion rather than just causation. That is, that we would be forced to act differently from how we would choose to act, under determinism. She thinks that it would be interesting to ask people what it would be like to be in a deterministic world. If they thought it would make us like puppets, or that we would be controlled by fate, then this would support the idea that they mistakenly think that in a deterministic universe people's actions would be coerced.

Another way that Nelkin thinks that people might mistake determinism to imply that an agent's mental states would not affect their behaviour, is by mistaking it to imply some form of epiphenomenalist reductionism. That is that our mental states have no effects on anything, including our behaviour. She notes that the idea that incompatibilist intuitions are based on the idea that agents are not objects, but that determinism implies that they are, has been suggested by Nagel (1979) (2007:256).

²⁴ Later she allows that emotion may also be involved, as I will discuss shortly.

Nelkin thinks that if people thought that living in a deterministic world would mean that we would be like robots it would suggest a mistake of this sort. She notes that Nahmias (2006) has found that people are more likely to attribute responsibility to agents in deterministic scenarios when the causation is described in psychological rather than neurological terms. This study employed the following scenarios and questions:

Neurological Scenario:

Imagine there is another universe similar to ours, in which there is a planet, named Erta, similar to ours in many ways. The landscape and life there look much like Earth, and there are advanced life forms (Ertans) who look, talk, and behave much like we do. However, the Ertans' science has advanced far beyond ours. Specifically, the Ertan neuroscientists have discovered exactly how Ertans' brains work. The neuroscientists have discovered that every single decision and action Ertans perform is completely caused by the particular chemical reactions and neurological processes occurring in their brain at the time, and that these chemical reactions and neurological processes in the brain are completely caused by earlier events involving their particular genetic makeup and physical environment. So, whenever Ertans act, their action is completely caused by the particular chemical reactions and neurological processes occurring in their brain at the time, and these brain processes are completely caused by earlier events that trace back to their particular genetic makeup and physical environment.

Psychological Scenario:

Imagine there is another universe similar to ours, in which there is a planet, named Erta, similar to ours in many ways. The landscape and life there look much like Earth, and there are advanced life forms (Ertans) who look, talk, and behave much like we do. However, the Ertans' science has advanced far beyond ours. Specifically, the Ertan psychologists have discovered exactly how Ertans' minds work. The psychologists have discovered that every single decision and action Ertans perform is completely caused by the particular thoughts, desires, and plans they have at the time, and that these thoughts, desires, and plans are completely caused by earlier events involving

their particular genetic makeup and upbringing. So, whenever Ertans act, their action is completely caused by the particular thoughts, desires, and plans they have at the time, and these thoughts, desires, and plans are completely caused by earlier events that trace back to their particular genetic makeup and upbringing.

(1) Now pretend that the scenario above is true and it accurately describes the Ertans. Assuming that is the case: Do you think that when the Ertans act, they can act of their own free will?

(2) Do you think that Ertans deserve to be given credit or blame for their actions? (2006:230-231)

Nahmias found that 18% (4/22) and 19% (4/21) of participants who received the neuro-scientific scenario judged that Ertans can act of their own free will and deserve credit or blame for their actions, respectively (2006:231-233). 72% (18/25) and 77% (17/22) of participants who received the psychological scenario made the same judgments, respectively.

Nelkin thinks that these results support the conclusion that people make a mistaken inference from determinism to epiphenomenalism. She thinks that there are more complex invariantist explanations of the data, that draw on ideas such as those above, and that these are more likely because of the inter-personal variation in the results.

In further support of the possibility that people mistake determinism for reductionism, Nelkin notes that, at a seminar, several psychologists objected to some of the experiments that are based on concrete deterministic scenarios (Nelkin, 2007:255). They thought that describing intentional actions in a deterministic world begs a key question, and is actually incoherent, because determinism rules out intentional actions. She thinks that psychologists would be better at judging when an experiment is problematic. It is not clear whether they mistake determinism for reductionism themselves, or think that most of the folk would. It is an interesting question whether psychologists who make this mistake themselves do so because of their training or pre-reflectively. It is certainly not obvious. But it may be plausible

that psychologists have a relatively good understanding of the folk's folk psychology. That may give them some authority in judging how the folk would understand concrete deterministic scenarios.

Nelkin thinks that Nichols and Knobe's experiments that control for concreteness suggest that emotional reactions effect people's responsibility judgments (2007:256-257). These experiments, as described in chapter two, used the questions:

As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?

As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?

64% of participants thought that it is possible that Bill is fully morally responsible for raping the stranger, and 23% thought that it is possible that Mark is fully morally responsible for cheating on his taxes. She thinks that this counts against the idea that the data can be explained solely in terms of concrete descriptions cancelling mistaken inferences from determinism. But she suggests an alternative explanation. She notes (like DKW) that NMNT found a high rate of responsibility attributions for their case involving bank robbery, even higher than Nichols and Knobe did for the case involving rape. As it is quite plausible that the rape case elicited stronger emotions than the bank-robbery case, this seems to count against emotion eliciting more responsibility judgments. I will argue in the final section of this chapter that there may be an important difference between Nichols and Knobe's and NMNT's descriptions of determinism. But it is unclear whether this is in terms of their tendency to cause people to infer that agents' mental states do not effect their actions. Seeing as this could explain the different responses to the bank-robbery and rape cases this difference does not give us any reason to reject emotion as an explanation in the cases that control for concreteness.

Nelkin also suggests that it could be seriousness rather than emotion that explains the difference in the responses to Nichols and Knobe's high and low-affect concrete cases. She thinks that NMNT's bank-robbery scenario is unlikely to be thought to be

more serious than the rape scenario, but this difference could be due to the different descriptions of determinism involved. Nichols and Knobe's may more strongly influence people to make the kinds of mistakes described above. Nelkin thinks that the contrasting of the deterministic and indeterministic universes in Nichols and Knobe's study might exacerbate this problem. She thinks that even if emotion does play a role in responsibility judgments it may be by complementing concreteness in cancelling out mistaken interpretations of determinism. Although emotion may be involved it might not fully explain the effect.

I think that seriousness may explain these results better, but I am not sure that we can clearly distinguish between seriousness and emotional salience. It seems that we would be serious about things that arouse our emotions, and vice versa. So I do not think that these points give us a good reason to reject the most obvious interpretation that emotion influences responsibility judgments. However, as I mentioned in the previous section, I agree with Nelkin that there is no good reason to think that emotion causes an error in these judgments. It could just as easily help people to see that determinism does not rule out someone's mental states causing their behaviour. It could do this by motivating people to think more carefully about it, or simply by priming concepts of intentional action. The research so far is not conclusive on this point.

Nelkin thinks that the points above may not explain all of people's intuitions about compatibilism, but that they show that there is not conclusive evidence that there is significant variation amongst them (2007:257). She is careful to note that she is only presenting a hypothesis, and that it may be undermined by further experimental research. She thinks that the complexity of the data raises avenues for further research into possible explanations, but that it may be difficult to separate the suggested confounding variables. It may be that recognising the difference between determinism and coercion or reductionism requires reflection upon the relevant concepts.

It may be worth pointing out here that Sommers (2010:205-206) has argued that NMNT's and Nichols and Knobe's studies described in section two are flawed in that they only test intuitions about incompatibilism and compatibilism directly. He

thinks that it would be more useful to test intuitions about the premises of philosophical arguments for those views, given that those arguments are supposed to do substantial work in drawing out the implications of determinism. This is a valid point. It is related to Nelkin's general argument in favour of testing the inferences that underlie people's judgments about compatibilism. Those inferences may include ones that philosophers suggest. For example, some people may infer that determinism rules out a person having alternative possibilities or being the ultimate source of their actions, and that one or both of these are required for responsibility, as incompatibilists argue. I agree that more research is needed here. My aim in this thesis is to consider one aspect of this research that is underway.

2. *Nahmias and Murray*

Nahmias and Murray (2010) have attempted to test the hypothesis Nelkin suggests.²⁵ They agree with her that people may only appear to have incompatibilist intuitions. They think that people do not usually see determinism as precluding responsibility until some explanation is given, that such explanations often lead them to think that determinism entails things that it does not, and that it is often only these things that they find incompatible with responsibility (2010:2-3). Many people may mistake determinism to imply what they call bypassing, which is that everything that happens will happen "*no matter what*", or that human psychology plays no causal role in human behaviour. More specifically this means that an agent's actions are caused by factors that bypass rational deliberation, conscious consideration of beliefs and desires, formation of higher-order volitions, and planning. Thus bypassing is the same general mistake that Nelkin is interested in, though Nahmias and Murray do not extrapolate it specifically in terms of coercion or epiphenomenalist reductionism. Bypassing is the mistaken inference that under determinism an agent's mental states have no effect on what they do.

²⁵ Page numbers for citations of Nahmias and Murray's paper are taken from the manuscript that was available prior to the recent publication cited. It is available at <http://www2.gsu.edu/~phlean/papers.html>

Nahmias and Murray note that both compatibilists and incompatibilists agree that determinism does not imply bypassing, and that bypassing rules out agents being responsible. It is generally accepted in the free will debate that, even if determinism is true, an agent's mental states can play a causal role in their actions. It could still be the case that if an agent had chosen to do something different then that choice would have caused them to do something different. An agent's behaviour, and other events, would not happen *no matter what*, but would depend on the prior conditions that determine that they happen. What compatibilists and incompatibilists disagree about is whether simply playing a causal role in their actions is enough to make an agent responsible for them. Incompatibilists claim that, as well as this, an agent's mental states or behaviour must not be determined by prior conditions at some point, or else that they be determined by the agent's self (see for example Kane, 2007:14-16, on self-forming actions). Nahmias and Murray thus describe their view as an error theory for incompatibilist intuitions, because it would show that they are based on a misunderstanding of the implications of determinism. It might not seem so simple to distinguish bypassing from the incompatibilist requirement for free will. I think that this distinction can be interpreted in different ways, one of which is problematic and one of which is not. But I will return to this point in section three.

They believe, in line with Nelkin, that a concrete competence model for responsibility judgments is correct (2010:8-9). They think that considering a specific agent intentionally performing a specific action in specific circumstances is more likely to make us consider the psychology of the agent, and see that bypassing is not implied by determinism. Like Nelkin, they note an earlier study in which most people responded that an agent was morally responsible in a deterministic world when their decisions were described as completely caused by their psychology, but not when they were described as completely caused by their neurology (Nahmias, et al., 2007).²⁶ They think that this is explained by people taking the latter to imply bypassing, but the former cancelling this by describing the agent's psychology.

²⁶ This is a later and more sophisticated study modelled on that reported in Nahmias (2006) which Nelkin refers to.

Nahmias and Murray ran an experiment to test these views (2010:9-10). It was based on Nichols and Knobe's abstract and concrete scenarios, and abstract and concrete versions of one of NMNT's scenarios. The main difference was that they included questions about bypassing as well as about free will and moral responsibility. Their specific hypotheses were as follows. (1) Most people who think that an agent in a deterministic scenario is not free and responsible would think that they were subject to bypassing. (2) Most people who think that such an agent is free and responsible would not think that they are subject to bypassing. (3) Attributions of responsibility and free will would be higher, and attributions of bypassing would be lower, in concrete scenarios than in abstract ones, especially in Nichols and Knobe's scenarios. (4) The same would hold for NMNT's abstract scenario compared to Nichols and Knobe's.

They randomly assigned one of the four scenarios to 436 undergraduate students from Georgia State University, Atlanta, using QuestionPro software to create and administer the survey online. The participants were first given an outline of the task, and then read the scenario. Nichols and Knobe's descriptions of deterministic and indeterministic universes, as described in chapter two, are:

Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries.

Now imagine a universe (Universe B) in which *almost* everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the

universe was exactly the same up until Mary made her decision, it *did not have to happen* that Mary would decide to have French Fries. She could have decided to have something different.

The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision *has to happen* the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision *does not have to happen* the way that it does.

The concrete condition based on this scenario included the following paragraph:

In Universe A, a man named Bill has become attracted to his secretary, and he decided 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. (2010:6)

The concrete scenario based on NMNT's description of determinism was as follows:

Imagine there is a universe (Universe C) that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature cause the exact same events for the entire history of the universe, so that every single time the universe is re-created, everything must happen the exact same way. For instance, in this universe a person named Jill decides to steal a necklace at a particular time and then steals it, and *every* time the universe is re-created, Jill decides to steal the necklace at that time and then steals it. (2010:5)

In the abstract case the final sentence was replaced with the following:

For instance, in this universe whenever a person decides to do something, *every* time the universe is re-created, that person decides to do the same thing at that time and then does it. (2010:10)

In order to replicate Nichols and Knobe's experiments, participants allocated these scenarios were first asked:

Which of these universes do you think is most like ours?

Universe A Universe B (2010:10)

Participants allocated NMNT's scenarios were instead asked:

Is it possible that our universe could be like Universe C, in that the same initial conditions and the same laws of nature cause the exact same events for the entire history of the universe?

Yes No (2010:11)

All participants were then asked to indicate on a 6-point rating scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree; scored as 1-6 respectively) to what extent they agreed with the relevant versions of the following statements about responsibility and bypassing.

Responsibility:

In Universe [A/C], it is possible for a person to be fully morally responsible for their actions.

[Bill/Jill] is fully morally responsible for [killing his wife and children / stealing the necklace].

In Universe [A/C], it is possible for a person to have free will.

It is possible for [Bill/Jill] to have free will.

In Universe [A/C], a person deserves to be blamed for the bad things they do.

[Bill/Jill] deserves to be blamed for [killing his wife and children / stealing the necklace.]

Bypassing:

In Universe [A/C], a person's decisions have no effect on what they end up being caused to do.

[Bill's/Jill's] decision to [kill his wife and children / steal the necklace] has no effect on what [he/she] ends up being caused to do.

In Universe [A/C], what a person wants has no effect on what they end up being caused to do.

What [Bill/Jill] wants has no effect on what [he/she] ends up being caused to do.

In Universe [A/C], what a person believes has no effect on what they end up being caused to do.

What [Bill/Jill] believes has no effect on what [he/she] ends up being caused to do.

In Universe [A/C], a person has no control over what they do.

[Bill/Jill] has no control over what [he/she] does.

Past Different:

In Universe A, everything that happens *has to* happen, even if what happened in the past had been different.

Suppose that in Universe C, a person named Jill decides to mow her lawn at a particular time and then does it. If Universe C were re-created with *different* initial conditions or *different* laws of nature, it is possible Jill would *not* mow her lawn at that time.

If Universe C were re-created with *different* initial conditions or *different* laws of nature, it is possible Jill would *not* steal the necklace at that time.

Bill *has* to kill his wife and children, even if what happened in the past had been different. (2010:11-12)²⁷

²⁷ The past different questions for NMNT's scenarios are only included in footnote 16, the first sentence of the abstract version only in footnote 24. I present them in full with the others in order to make comparisons easier.

The question about moral responsibility was asked first in order to replicate Nichols and Knobe's study. The other questions were given in random order to control for order effects. The participants also answered two manipulation checks, which are not reported, and some demographic questions. 42% of the participants were male and 58% were female. 187 of the 436 subjects were excluded from analysis for failing to complete the survey, failing one of the manipulation checks, or completing the survey too quickly.

The results were analysed by averaging the responses to the responsibility questions, and the bypassing questions excluding the past different ones, to get composite responsibility and bypassing scores. People who scored less than 3.5 on the composite responsibility score were classed as 'apparent incompatibilists', and people who scored more than 3.5 on the composite bypassing score were classed as 'bypassers'. 20 subjects who scored exactly 3.5 on the composite bypassing score were excluded.

There was a strong inverse correlation between the composite bypassing and responsibility scores. Generally, the higher the bypassing score the lower the responsibility score. In the abstract cases the scenario had no statistically significant effect on the composite responsibility scores over and above the effect it had in virtue of effecting composite bypassing scores, which was statistically significant (2010:13-14). Most people who were classed as apparent incompatibilists were also classed as bypassers. 80% and 67% in Nichols and Knobe's abstract and concrete cases, respectively, and 85% and 100% in NMNT's abstract and concrete cases, respectively (2010:23). Thus Nahmias and Murray's first hypothesis [most people who think that an agent in a deterministic scenario is not free and responsible would think that they were subject to bypassing] was apparently proved correct. Most people who were not classed as apparent incompatibilists were not classed as bypassers. 72% and 75% in Nichols and Knobe's abstract and concrete cases, respectively, and 85% and 77% in NMNT's abstract and concrete cases, respectively (2010:23). Thus Nahmias and Murray's second hypothesis [most people who think that such an agent is free and responsible would not think that they are subject to bypassing] was apparently proved correct.

NMNT's scenarios received statistically significantly higher composite ratings for responsibility than Nichols and Knobe's, and the concrete scenarios received statistically significantly higher composite ratings for responsibility than the abstract ones. 73% (of 77) and 29% (of 56) of participants were classed as apparent incompatibilists in Nichols and Knobe's abstract and concrete cases, respectively, whereas 48% (of 56) and 18% (of 60) were classed as apparent incompatibilists in NMNT's abstract and concrete cases, respectively (2010:22). The concrete scenarios also received statistically significantly lower composite ratings for bypassing than the abstract ones. However, only NMNT's abstract scenario, and not their concrete one, received statistically significantly lower composite ratings for bypassing than Nichols and Knobe's corresponding scenarios. 66% and 49% of participants were classed as bypassers in Nichols and Knobe's and NMNT's abstract cases, respectively, whereas 37% and 35% were classed as bypassers in Nichols and Knobe's and NMNT's concrete cases, respectively (2010:22). Thus Nahmias and Murray's third and fourth hypotheses [attributions of responsibility and free will would be higher, and attributions of bypassing would be lower, in concrete scenarios than in abstract ones, especially in Nichols and Knobe's scenarios; and the same would hold for NMNT's abstract scenario compared to Nichols and Knobe's, respectively] were apparently proved correct.

Nahmias and Murray think the most plausible interpretation of these results is that when people take determinism to imply bypassing this causes them to think that it rules out responsibility (2010:15). Also when people do not infer bypassing they do not see any conflict between them. But as determinism does not entail that our decisions, beliefs, and desires have no effect on our actions, or that we have no control over our actions, they think that their error theory for incompatibilist intuitions is confirmed.

They think that the higher rate of bypassers in the abstract cases compared to the concrete ones may be due to the description of concrete agents and actions causing people to consider the effectiveness of the person's psychology (2010:15). They also think that concrete cases are more likely to provoke reliable intuitions because they include more information about a scenario (2010:8). They agree with Nichols and Knobe that the emotion provoked in their concrete case (Bill burning down his

family home) may bias people's responses, but they do not think that the responses are more reliable in the abstract case (2010:16). Rather they think that emotion bias may simply be making up for a strong suggestion of bypassing in the abstract case. They think that the performance error model cannot account for the compatibilist majorities in NMNT's cases involving positive and neutral actions since these did not involve high affect. These cases, as described in chapter two, involved an agent saving a child from a burning building, and going jogging, respectively. Nahmias and Murray think that these cases are comparable in affect to Nichols and Knobe's low-affect case (cheating on taxes). But since similar proportions of participants judged that the agents performed those actions of their own free will, as did for the case in which Jeremy robs a bank (68%, 79%, and 76%, respectively) they think that affect cannot account for them.

Nahmias and Murray think that Nichols and Knobe's description of determinism suggests bypassing especially strongly because it may be interpreted as saying that what happens does not depend on the past (2010:7-8). This criticism has been made before (for example in Nahmias, 2006:221-224, and Feltz, et al., 2009:13-16, and of a similar experiment in Turner, et al., 2006:605-608). There are two parts of it that may lead to this interpretation. Firstly, the description of the deterministic universe says that given the past each decision *has to* happen the way that it does, rather than that given the past each decision *will* happen the way that it does. Nahmias and Murray point out that this is ambiguous in terms of the scope of the necessity involved. It might be interpreted as given the past, necessarily what happens will happen as it does, rather than as necessarily, given the past what happens will happen as it does. The former interpretation may lead people to think that even if the past were different the same things would happen. Secondly, the final paragraph of the vignette says that in the indeterministic universe, in contrast to the deterministic one, decisions do not have to happen as they do. It should say that in contrast *given the past* decisions do not have to happen as they do. This is a problem because in contrasting the two universes the participants may infer that in the deterministic universe decisions will happen the way they do regardless of the past, rather than only given the past.

This is a problem partly because it rules out conditional abilities to do otherwise. Again, a conditional analysis of ability says that being able to do other than what you do means that you would do something else if some prior condition was different from the way it was. The prior condition is usually an intentional factor like desiring or choosing. Now determinism is clearly compatible with being able to do otherwise in this sense. But the interpretation above implies that even if an agent had chosen to do something else they would not have, because what happens does not depend on the past. So compatibilists who require a conditional ability to do otherwise for free will would not attribute responsibility if they understood the description of the determinist universe in this way. Thus their responses would be misinterpreted as representing incompatibilist intuitions. This is an even bigger problem because it may lead people to infer bypassing. It implies that an agent's mental states have no effect on their actions. Even compatibilists in the Frankfurt tradition, who do not require alternate possibilities (abilities to do otherwise) of any kind for responsibility, require that one's actions are causally connected to one's mental states.

Nahmias and Murray think that this problem is supported by the fact that 48% of participants in Nichols and Knobe's abstract case responded that everything that happens has to happen the way it does even if the past had been different (2010:15). There was a statistically significant correlation between these responses and the composite responsibility score. In NMNT's abstract case only 7% of participants responded that if the universe described were recreated with different initial conditions or different laws of nature it would not be possible that Jill would not mow her lawn. They think that this is evidence for their claim that Nichols and Knobe's description of determinism is problematic in using the phrase "has to happen" and in omitting the phrase "given the past" in the final paragraph. They think that this, along with the other problems raised earlier, explains the difference in the proportion of bypassers in the abstract cases. They think that the high affect in Nichols and Knobe's concrete case lowers the rate of bypassers there.

They thus take low affect concrete cases to be most reliable (2010:16). They note that NMNT's concrete case was most like this in that the idea of someone stealing a necklace is less likely to arouse strong emotions than that of someone murdering his

family. In that case every apparently incompatibilist response was correlated with bypassing and every non-bypassing response was correlated with an apparently compatibilist response. So as well as eliciting the fewest number of bypassing responses it exhibited the least variability among those who did not make that mistake. They think further research should explore bypassing and its relation to responsibility judgments in other concrete cases that have already been studied, including those involving positive and neutral actions, and should try to come up with high-affect abstract cases. Presumably they expect that such research would find further evidence that in concrete cases most people do not infer bypassing, and that these people would also attribute responsibility to the agent involved. They also seem to imply that in high-affect abstract cases most people would infer bypassing, and that these people would judge that it is not possible for people to be responsible for their actions in the world described.

Nahmias and Murray also criticize Nichols and Knobe's descriptions of deterministic and indeterministic worlds for mentioning humans in the description of universe B but not universe A (2010:7). They think that this may prime some subjects to think that universe B is more like ours. They think that this may be exacerbated in the abstract condition, because in the concrete scenario the description of Bill's behaviour suggests that he is human. They think that this may explain some of the difference in judgments about responsibility because whether a scenario involves alternate or real-world universes has been shown to affect them (they cite Nahmias, et al., 2007, and Nichols, et al., 2008).

They consider one objection to their conclusion that most people do not have genuine incompatibilist intuitions (2010:16-17). It is that upon understanding that determinism does not imply bypassing people would or should think that it threatens responsibility, and that those who do not probably do not understand the determinism in the scenarios. This implies that the apparently compatibilist intuitions are not genuine, perhaps because they do not see that determinism rules out alternate possibilities or being the ultimate source of one's actions. They asked two questions that they think are relevant to this. In the NMNT scenarios they asked whether it is possible, if the universe were recreated with the same initial conditions and laws of nature, for Jill not to mow her lawn, or steal the necklace. They think

that the objection above would predict that people who thought that this is possible, and thus that the determinism involved does not rule out alternate possibilities, would be more likely to attribute responsibility. But there was no significant correlation in the abstract case (mowing the lawn), and only a marginally significant one in the concrete case (stealing the necklace).

They think that further research should try and check whether people take determinism to rule out alternate possibilities or being the ultimate source of their actions, and whether they take these things to be required for responsibility (thus they agree with Sommers, 2010). But they think that it will be difficult to ask such questions without explaining the concepts clearly which may verge on intuition coaching. They note that there is a trade off between making sure that participants understand the relevant concepts, and seeing what they think without any philosophical training, due to the concern that it may introduce bias. Since philosophical intuitions exhibit conflict, as well as philosopher's claims about folk intuitions, we should try and see what the folk intuitions and their sources are, and whether they are reliable.

Like Nelkin, Nahmias and Murray do not think that bypass necessarily explains all of the conflicting intuitions discovered in the studies above. They think that there are other problems with Nichols and Knobe's study. One is that the phrase "fully morally responsible" used in the questions is somewhat technical sounding, ambiguous, and likely to be interpreted differently by different people (2010:7). The term "fully" may be contrasted with being partially, or in some lesser sense, responsible. One notorious ambiguity in concepts of moral responsibility is between forward-looking deterrent concepts and backward-looking retributive concepts. The former refers to our holding someone morally responsible because by doing so we can cause them, or others, to behave differently in the future. The latter refers to our thinking that someone is responsible for their behaviour and thus deserves to be punished just because they have behaved immorally already, regardless of the consequences of punishing them. They think that "fully morally responsible" lends itself to the interpretation that it is about whether the agent should be punished, and that subjects may think they should be punished even if they are not fully morally responsible. This interpretation may not be primed in the abstract case because no

specific act is mentioned for which the agent could be punished, partly explaining the difference in responses to abstract and concrete cases. They think that asking about whether an agent deserves blame or praise is more likely to prompt the notion that is relevant to the debate. This point relates to the ambiguity thesis which I will discuss in section four.

They conclude that while some professional philosophers obviously do have genuine incompatibilist intuitions, the evidence so far suggests that most of the folk do not (2010:18). Evidence that the folk do find incompatibilism intuitive when they do not mistake it to imply bypassing is required to refute this. Thus they claim to have provided support for Nelkin's hypothesis that most folk intuitions about compatibilism are invariant and compatibilist. As I have noted there are still questions about the inferences that underlie people's judgments in these studies, including whether they include consideration of the inferences involved in philosophical arguments about free will. While these have yet to be studied empirically, the data that is currently available is also subject to the challenges raised by Nelkin, Nahmias, and Murray.

3. Problems with the Experiment

I think that this study has a plausible argument supporting it, and important goals in trying to understand how people interpret deterministic scenarios, and what it is about them that influences their responsibility judgments. The strategy of checking for errors underlying folk intuitions is what is required to either consolidate the results found so far or move our understanding forward. But I think that Nahmias and Murray overestimate the strength of their results. The conflicting responses are underemphasized in their analysis, and several of the experimental questions are subject to alternate interpretations which undermine their conclusions.

If the bypass questions were interpreted as Nahmias and Murray expected then there would be evidence for their error theory for many incompatibilist intuitions. But they underplay the inconsistencies in their results. There were arguably significant minorities who had incompatibilist intuitions but who were not bypassers: 15% in Nichols and Knobe's abstract case, 10% in their concrete case, and 7% in NMNT's

abstract case. It seems that these participants ruled out responsibility without inferring bypassing. The most obvious explanation is that they thought that determinism ruled out responsibility. There may of course be some other explanation of these judgments. But this research has to consider one explanation at a time. It is generally accepted that people do judge that people are responsible for their actions much of the time. The most obvious difference between the scenarios employed in this experiment and the real world is the specification of determinism. When people deny that an agent is or could be responsible in these scenarios, this seems to show that they find incompatibilism intuitive. Of course this cannot be concluded with absolute certainty while there are other plausible explanations. But no one is claiming that it can. This is an area of research in which there is still much work to be done. What I am trying to do is argue for the best, and therefore tentative, interpretation of that research in its current state.

There were also even greater minorities of participants in Nahmias and Murray's experiment who had compatibilist intuitions but who were bypassers: 8% in each of the abstract cases, 18% in Nichols and Knobe's concrete case, and 19% in NMNT's concrete case. If the questions were interpreted as Nahmias and Murray intended them then bypassing did not undermine responsibility for these participants. This seems to imply that some of the participants who had incompatibilist intuitions and were bypassers, may not have taken bypassing to rule out responsibility either. It seems that for these participants, as well, determinism was enough to rule out responsibility. Whether they make up a significant group depends on the questions about inter-personal variation in intuitions that I raised in chapter two, but which I will not attempt to resolve in this thesis.

But I think that the fact that many participants who were classed as bypassers still judged that agents were responsible suggests that the questions about bypassing were not always interpreted in the way that Nahmias and Murray intended. If they had been then those participants' responses would indicate that they think that an agent can be responsible for an action that was not caused by their mental states, and that they had no control over. It seems implausible that people would think this. Recall that bypassing is the idea that an agent's mental states have no effect on their behavior. Nahmias and Murray classed a participant as a bypasser if they scored

above the mid-point on average in the questions about bypassing. I think that there is a different way that participants may have interpreted the bypassing questions that better explains the results. Specifically I think that participants may have interpreted the causation involved in those questions in different ways.

Nahmias and Murray intended that causation in terms of something (A) simply playing a causal role in, or being part of the causal history of, the occurrence of something else (B). I will call this simple causation. The idea of bypassing, which they were trying to isolate, is that an agent's mental states do not play a causal role in, and are not part of the causal history of, their actions. It is an error to think that this is implied by determinism. But another kind of causation is when something (A) is a simple cause of something else (B), but when there are also no prior conditions that determine that (A) will occur. I will call this undetermined causation. Clearly, if determinism is true then there are no undetermined causes (except the first event or initial conditions, if there are any) because for every event there are prior conditions that determine that it will occur. Thus it is no error to think that an agent's mental states cannot be undetermined causes of their behavior if determinism is true. I think that many participants may have interpreted the bypassing questions in terms of undetermined causation. That is, for example, they may have interpreted the item "In Universe A, a person's decisions have no effect on what they end up being caused to do." as 'In Universe A, a person's decisions have no effect on what they end up being caused to do, or they have an effect on what they end up being caused to do but there were prior conditions that determined that those decisions would occur as they did.' Thus by responding as what Nahmias and Murray classified as bypassers, they may not have been judging that bypassing was implied by the deterministic scenarios. They may simply have recognized that in the scenarios an agent's mental states are determined by prior conditions. They are not undetermined causes. While they *are* part of the causal history of our actions, their mental states play *a* role in what happens, they are not the ultimate source of them, and do not have sufficient control to be responsible.

It may seem implausible to philosophers, who likely have much clearer concepts of causation, that anyone would interpret the bypassing questions in terms of undetermined causation. To see how it could be natural (though not necessarily

justified) for the folk to do so, consider the following. If something (A) causes something else (B), and (B) causes something else (C), then it seems that (A), and not (B), is in a sense the ‘real’ cause of (C). What the term ‘real’ seems to be getting at is that (A) is the original cause of (C). It is the start of the chain of causation that led to (C). In other words, there was nothing that caused (A) to occur. If (Z) had caused (A) to occur then it would seem that (Z) was, in a sense, the ‘real’ cause of (C). But if nothing causes something to happen then there are no prior conditions that causally determine that it would occur. It would be an undetermined event. So what ‘real’ is really getting at is ‘undetermined’. Thus it seems that the concept of an event that causes something, but for which there are no prior conditions that determine that it will occur, is quite naturally separated from that of an event that simply causes something. The concepts of simple and undetermined causes are plausibly common among the folk. Of course this would require experimental testing to confirm. But here I am simply describing a plausible alternative explanation of the results to show that Nahmias and Murray’s conclusion is not warranted on the basis of them alone.

Given that undetermined causation is a concept that is readily accessible to the folk, there are several aspects of this experiment that may have made it more likely that the participants interpreted the bypassing questions in terms of it. Given the evidence described in chapter two, that people do not ordinarily think that determinism is true for human actions, this aspect of the scenarios probably stands out as the point of interest in the experiment. It is the most obvious way that they differ from the participants’ understanding of the actual world. This may have caused them, consciously or otherwise, to assume that the causation of interest in the bypassing questions was undetermined causation. The phrasing of the questions may have amplified this effect. Consider, for example, “In Universe [A/C], a person’s decisions have no effect on what they end up being caused to do.” The fact that the scope of the causation of the mental states in the questions is “what they end up being caused to do” rather than simply ‘what they do’ seems likely to reinforce the message that it is the causation in the scenarios that is important. The fact that human behavior is deterministically caused, again, may suggest that the questions are about whether agents are determined to do what they do, rather than whether they were simply part of the causal history of their actions. Phrasing the questions

negatively may also have suggested that they were about undetermined causation. It is an unnatural way to phrase a question. As anyone who understood the scenarios would see that simple causation is not ruled out, but undetermined causation is, they may have taken this as a signal that the questions were about the latter.

If all of the participants had interpreted the bypassing questions in terms of undetermined causation then all of those who were not classed as bypassers would have misunderstood the scenarios. They would have judged that agent's mental states do have undetermined effects over their behavior, when clearly the deterministic nature of the scenarios rules this out. But there is a good reason to think that some people may have interpreted the questions in terms of simple causation and others in terms of undetermined causation. That is that different kinds of causation may be more important to different people. Simple causation would be more important to compatibilists, and undetermined causation would be more important to incompatibilists.

Incompatibilists require that determinism be false for an agent to be responsible for their actions. But they do not simply require that there be some events in the world that are indeterministic in order for people to be responsible for what they do. As I stressed in chapter one, indeterminism is not synonymous with, and does not ensure the existence of, free will. After all the mere falsity of determinism would not plausibly empower anyone to be responsible. Indeterministic events that have nothing to do with an agent's behavior are irrelevant to their being responsible for something. What incompatibilists require is that some of an agent's behavior not be determined by prior conditions. Of course it is usually also required that an agent cause their actions. It is hard to see how someone could be responsible for something that they did not cause. As Nahmias and Murray point out, both compatibilists and incompatibilists agree that an agent's mental states must affect their behavior in order for them to be responsible for it, and this is not ruled out by determinism. Thus an agent must play a certain causal role in their actions, to be responsible for them on an incompatibilist view. They must be an undetermined cause, one for which there are no prior conditions that determined that it would occur as it did. Compatibilists also require that an agent cause their actions in order that they be responsible for them. But they do not require that that causation not be

determined by prior conditions. It is enough that they be a simple cause of them. Of course there are certain ways that this simple causation must be on compatibilist views. They vary, but for example a common requirement is that it not involve coercion. This, then, is a fundamental difference between compatibilists and incompatibilists. Compatibilists require that an agent be a simple cause of their actions in order to be responsible for them, while incompatibilists require that they be an undetermined cause of them. This is not to say that people have different concepts of responsibility, which is the idea behind an ambiguity thesis, as I will discuss later. Rather the idea is that people require different things in order to be responsible in the same way. As responsibility is a very important thing to people the relevant kind of causation they require is also likely to be important to them. It is thus likely that they think about causation in different ways. If some participants have compatibilist intuitions and some have incompatibilist intuitions, then they are likely to have interpreted the bypassing questions in terms of different kinds of causation. In the same way, if participants have compatibilist intuitions in some contexts and incompatibilist intuitions in others, it is likely that they interpreted the bypassing questions in terms of different kinds of causation in different contexts. Thus those questions might not easily distinguish whether variation in folk intuitions exists, either inter-personally or intra-personally.

Where the bypassing questions asked whether the agent's decisions, beliefs, or wants have no effect on what they do, participants may have interpreted them as asking whether they have no undetermined, or no simple effect on what they do. The questions about control could be subject to a similar effect. Folk incompatibilists probably think that being determined by prior conditions removes real, or responsibility-conferring, control from an agent. It would then be the prior conditions that control their actions. Participants may thus have interpreted the questions as being about whether the agent has undetermined control, or whether they have simple control. Undetermined control seems to be the kind incompatibilists intend when they require that an agent have alternative possibilities to, or be the ultimate source of, their actions, in the ways precluded by determinism.

The fact that the participants were also asked about moral responsibility makes it more likely that they would have interpreted the bypass questions differently. This

could have subconsciously primed participants to engage the concept of causation or control that they take to be required for someone to be morally responsible. This interpretation could even have been selected at a conscious level. The participants may have consciously inferred that the bypass questions were supposed to be about the kind of causation or control that is required for responsibility. In that case they may have recognized that a strict interpretation would be about whether an agent is a part of the causal history of an action, but deliberately interpreted it as being about what matters for responsibility. And if they were incompatibilists then that would be about undetermined causation and control. This effect could well have been amplified by the fact that the first question that the participants were asked was always about moral responsibility. Recall that while the other questions were presented in random order to control for order effects, the question about responsibility was always presented first, in order to replicate Nichols and Knobe's study.

This can explain why there was a strong correlation between having incompatibilist intuitions and being classed as a bypasser, without there being any underlying error. Incompatibilists could have interpreted the bypassing questions in terms of undetermined causation and control, and correctly responded that agents do not have them in the scenarios. Compatibilists could have interpreted them in terms of simple causation and control, and correctly responded that agents do have them in the scenarios. Thus the bypassing questions in this experiment do not add much extra weight to the argument that incompatibilist folk intuitions are based on a mistaken belief that determinism implies bypassing. There is a plausible alternative explanation that does not involve the participants making a mistake about the implications of determinism.

Nahmias and Murray think that the responses to the questions about what is possible if the past were different in the abstract cases also constitute evidence that some incompatibilist intuitions are based on bypassing. Recall that the past different questions for the abstract cases (Nichols and Knobe's, and NMNT's, respectively) were:

In Universe A, everything that happens *has to* happen, even if what happened in the past had been different.

Suppose that in Universe C, a person named Jill decides to mow her lawn at a particular time and then does it. If Universe C were re-created with *different* initial conditions or *different* laws of nature, it is possible Jill would *not* mow her lawn at that time.

Nahmias and Murray think that the responses to the past different questions may explain the responses to the bypassing questions in these scenarios, which in turn explain the responsibility judgments. They think that the fact that many more participants agreed with the statement for Nichols and Knobe's scenario than did for NMNT's shows that it suggests that what happens does not depend on the past. This would imply that an agent's behavior does not depend on their prior mental states, which is bypassing. Although the bypassing questions are flawed, as described above, the past different questions may provide independent evidence that some participants did infer bypassing. But again I think that there is ambiguity in the past different questions as well, that undermines the conclusion.

The problem with the past different questions is that they were asked in different ways for the two abstract cases. In Nichols and Knobe's they asked whether everything that happens has to happen, even if the past had been different. In NMNT's they asked whether, if the past had been different, it would be possible that a specific agent would not perform a specific action that that she actually did. I think the latter is much clearer. In the former some of the participants who agreed might have thought that the question concerned whatever would happen if the past were different, rather than the things which actually happened. In that case by agreeing they could have meant that different things would happen if the past were different, but that those things would have to happen based on the different past. This does not imply that they made a bypass error because it allows that different actions could be performed based on different prior mental states. In the question about the past being different for NMNT's case it is clear that the subject is an event that actually happened, rather than something that would have happened if the past had been different.

There are also plausible alternative explanations for the different responsibility scores in the different scenarios, other than their being caused by participants inferring bypassing. Recall that there were significantly lower rates of responsibility attributions in the concrete cases compared to the abstract ones, and in NMNT's cases compared to Nichols and Knobe's. The concrete cases are possibly subject to the emotion effect discussed. Nahmias and Murray may be right that an emotion-based model alone cannot explain the fact that similar rates of responsibility attributions were found in NMNT's experiments involving neutral and praiseworthy actions (going jogging and saving a child). These cases may not have provoked an emotional response greater than Nichols and Knobe's low-affect concrete case (cheating on taxes) with which they supported their performance error model. However this does not resolve the issue one way or the other. Nahmias and Murray's concreteness model cannot explain the result in Nichols and Knobe's case without independent evidence of its tendency to imply bypassing or some other error. I have now discussed this conflict at several points. To be clear, I am not arguing in favor of the performance error model, or for any emotion-based model. Clearly there is a problem in interpreting these results. I think that there are plausible arguments for several models, as Nichols and Knobe originally asserted. More empirical research will be required to find a satisfying explanation. Until such research is carried out, and stands up to criticism, there are several possible explanations. But as I argued in chapter two, the existence of non-error-based variation in folk intuitions does not obviously hinge on either concreteness or emotion explaining these results. And while there may be other explanations for them which do show that the variation is due to an error, they are not obvious, and they need to be discussed and tested one at a time.

It is interesting that the new abstract version of NMNT's scenario received lower rates of responsibility attribution than Nichols and Knobe's. An alternate explanation for this is one that Nahmias himself (and others) has suggested, which I discussed in section two. That is that some participants are understanding Nichols and Knobe's description of determinism, but not NMNT's, as ruling out agents having conditional alternate possibilities. An agent has conditional alternate possibilities if they would have acted differently if their mental states had been

different prior to their action, and their mental states could have been different in such a way. Not having conditional possibilities does not imply bypass, because it could be that an agent's mental states are part of the causal history of their actions but that those mental states could not have been different. As mentioned in chapter two, Nichols and Knobe think that most compatibilist philosophers these days are persuaded by Frankfurt cases that neither conditional nor unconditional alternate possibilities are required for moral responsibility. They intended their description of determinism to test for this kind of compatibilism, which is why they used the phrase "has to happen" to rule out alternate possibilities. Nahmias and Murray criticise this strategy, as mentioned in the previous section, in part by noting that compatibilists may require conditional alternate possibilities for responsibility and determinism does not rule these out. Nahmias notes in an earlier paper that many philosophers still require conditional alternate possibilities. NMNT developed their descriptions of determinism to test for compatibilism in general rather than only for Frankfurt style compatibilists. Now it seems plausible that some folk require conditional alternate possibilities for responsibility and some do not require alternate possibilities at all, given that this is true of philosophers. In that case if the participants understood Nichols and Knobe's and NMNT's scenarios as they were intended, then those who were compatibilists and require conditional alternate possibilities for responsibility would tend to attribute responsibility in NMNT's ones but not in Nichols and Knobe's. This could explain the different rates of responsibility attributions between them without it being due to more participants inferring bypassing from Nichols and Knobe's description than NMNT's.

It is plausible that people would have interpreted Nichols and Knobe's description of determinism as ruling out conditional alternate possibilities but not implying bypassing. That is because it says that everything that happens from the beginning of the universe has to happen given the past. Rather than taking this to mean that what happens has to happen regardless of the past, some people may have taken it to mean that the recent past, including people's mental states, could not have been different, given the earlier past. Thus conditional alternate possibilities would be ruled out because although people would act differently if they had chosen to, they could not have chosen differently. The difference in responsibility judgments between them could be explained by some of the participants inferring that

conditional alternate possibilities are ruled out in the former but not the latter. This would support another criticism of Nichols and Knobe's experiments. It would also mean that the minority of participants who find compatibilism intuitive in both abstract and concrete cases is greater than those experiments suggested. This would show that there was even more inter-personal variation, making the questions raised in chapter two even more pressing. But this is only possible, and we cannot tell from the data whether it is this, a tendency to cause bypass error, or both that are causing the different rates of responsibility attributions. Nahmias and Murray are right that further research is required here.

There is another possible explanation for the difference in the responsibility judgments between Nichols and Knobe's and NMNT's abstract scenarios. That is that Nichols and Knobe's description of determinism is just clearer, so that more of the incompatibilist participants deny that agents could be responsible in it. After all it says that things have to happen *given the past*, and *because things are completely caused by what happened before them*. This more obviously implies determinism than it does that an agent's past mental states have no causal role in their actions, or could not have been different. The phrase "has to happen" might not be interpreted in terms of logical necessity by the folk. It might be interpreted in terms of what might be called causal necessity. That is not just that given the past the events that happen will happen, but that they are causally connected, so that the past causes them to happen. There is arguably a difference between events simply following from the past and being caused by them. While Nichols and Knobe's description is ambiguous between logical and causal interpretations it is not clear that folk would take the problematic one. So this description may well have been interpreted in terms of determinism, and perhaps even more so than NMNT's. It is longer than NMNT's description, and involves comparison with an indeterministic universe that could make it clearer. The fact that the subjects whose responses are reported correctly answered two manipulation checks speaks against there being a significant difference in terms of correctly inferring determinism from the scenarios. But this is mitigated by the fact that the manipulation checks themselves are not reported. Given the flaws in the other questions in this study it is not clear that we should have a great deal of confidence in them.

Again, I am not trying to establish any particular explanation of the complex patterns of responsibility judgments in this experiment. I think that much more empirical research will be required to do so. What I am arguing is that there are explanations, other than bypassing, which could explain those patterns. Thus, this study has not provided significant evidence in favor of the hypothesis that incompatibilist intuitions are based on erroneously inferring bypassing from determinism. The questions that it has employed are not clear enough to do so. As Nelkin warned, the subtle concepts involved may be extremely difficult to separate. While experimental studies are the only way to resolve these issues, they have to be designed very carefully, as the problems with this one show.

4. The Ambiguity Thesis

Another possibility for explaining away the apparent variation in folk intuitions discovered in the studies described in chapter two is there being more than one concept of responsibility or free will that the participants were employing. I will call this an ‘ambiguity thesis’ because the idea is that terms like ‘free will’ and ‘responsibility’ are ambiguous in that they may refer to multiple concepts. If some kind of conceptual confusion were taking place it would seem that there might not be variation in folk intuitions about the same concept of responsibility. Rather it could be that there is variation in when the folk employ different concepts of responsibility, so that for each concept the folk might actually all have the same intuitions about it, and have the same intuitions in different cases. As NMNT (2006:46) describe one possibility “when people use ‘free will’ in contexts and ways intimately tied up with practices of moral responsibility, sometimes it expresses one content, compatible with determinism, and other times - notably, in philosophical discussions when the criteria of applicability become more stringent - it expresses another content, incompatible with determinism.”

There are different possibilities for how an ambiguity thesis might pan out. It could be, for example, that the participants sometimes responded that the agent was, or that agents could be, responsible or free, and meant that they were, or could be, legitimately called to moral improvement. That is, that they would or could respond

to praise or blame by changing their behavior in the future. At other times their responses may have meant that they were or could be deserving of punishment or reward. That is, that it would, or could, be morally right to punish or reward them for their actions, regardless of the effects it would have on their behavior. This is a common distinction in the literature. Nahmias and Murray (as mentioned) call these concepts ‘forward-looking’ or ‘deterrent’ and ‘backward-looking’ or ‘retributive’ concepts respectively. Pereboom (2007:86-87) calls them ‘legitimately called to moral improvement’ and ‘basic desert’ respectively. Vargas (2006:246-247) calls them ‘pragmatic’ and ‘metaphysical’ respectively. The consensus seems to be that it is only the latter that is the subject of the free will debate, and that the former is clearly compatible with determinism. Thus it seems most plausible that if this kind of ambiguity exists then the apparently compatibilist intuitions found in the studies may be about the former concept, and the apparently incompatibilist intuitions about the latter.

What this would mean would be a subject for debate. Pereboom has argued that we should stop attributing responsibility of the basic desert kind to people, and that this would have positive effects on individuals and society (2007:123-124). Vargas, by contrast, has argued that basic desert may not be the most fundamental concept of desert people use concerning responsibility. He thinks that even if it is we may be able to revise it in non-basic ways, and that we should do so if it would promote valuable kinds of agency (2007:155 and 210-211).

Feltz et al. (2009:8-10) sought to test for an ambiguity thesis (what they call the ‘multiple concepts view’) empirically. They used Nichols and Knobe’s high and low-affect scenarios (rape and cheating on taxes) but presented both to each of 52 participants. They found that 67% gave incompatibilist answers to both, and only 8% gave mixed answers (one compatibilist and one incompatibilist answer). They think that this *prima facie* supports an ambiguity thesis, though they do not describe the content of the different concepts that they think exist. But, they do not give any reasons to think that their results are due to an ambiguity thesis, rather than there just being inter-personal variation in folk intuitions about the same concept. Perhaps they do not discuss this result much because they go on to argue that there are

serious problems with all the experiments on folk intuitions about free will that have been run so far, including this one (2009:19).

It would take further experimental research to show that an ambiguity thesis is correct, and the particular form that it takes. The implications of an ambiguity thesis would seem to hinge on the content of the particular concepts of free will and responsibility, as well, of course, as which ones are intuitively compatible or incompatible with determinism. This is essentially the point raised in chapter one, that for any particular concept of free will, and any particular concept of determinism, there is a separate compatibility question. So this is an interesting area for further research, and one that could potentially explain away variation in intuitions about free will. I will not offer arguments for or against an ambiguity thesis being true. My goal in this thesis is to explain the current state of one interesting aspect of experimental philosophy on free will, and to criticize one recent experiment within it. My point is that there at least appears to be variation in folk intuitions about free will, that if there is it raises interesting and novel questions, and that Nahmias and Murray's study does not show that there is not.

Conclusion

In this thesis I have attempted to show that there are significant problems with Nahmias and Murray's experimental design and analysis that undermine its support for their error-theory about incompatibilist intuitions. Their results are not as one-sided as they imply, and the key experimental questions are prone to misinterpretation. Thus the unifying theory for the intuitions discovered in experimental studies on free will and determinism, that they and Nelkin propose, does not find significant empirical support there. As that unifying theory remains a relatively plausible hypothesis, further research is required to settle the issue. For now though, DKW's claim that some experimental studies suggest intra-personal variation in folk intuitions regarding compatibilism remains plausible. It also remains plausible that folk intuitions about compatibilism are subject to inter-personal variation, as I have emphasized.

The combination of conservatism and invariantism that is so widely exhibited is at least questionable on this basis. So are the assumptions that the free will debate applies to everyone in the same way, and that intuitions about it can be seen to support philosophical theories due to their being widely-shared. Further research will be required to clarify the position of these assumptions, and the implications of rejecting them, if that is required. Both empirical and conceptual work is relevant. Empirical studies informed by hypotheses of the Nelkin/Nahmias/Murray kind could show that variantism and the appearance of significant groups with conflicting folk intuitions is only superficial. If this is not the case it is likely that further empirical work will be required to assess the extent of variation in folk intuitions about free will, and the particular forms that it takes. This information will help to settle the conceptual issue of whether variantism or conservatism, or the requirement that intuitions be generally shared in order to support theories, should be rejected.

Experimental research is flourishing in this area, and I hope that it is clear that it is well-motivated. While it is a difficult project, it may become easier as more evidence comes in, and as philosophers gain more experience with it. It is not surprising that folk intuitions should not be one-sided in a debate that philosophers have found so intractable. Recognizing this fact, if it is confirmed, may help clear

the way to a more complex but more accurate understanding of ordinary thinking about responsibility.

Bibliography

- Alexander, Joshua, and Jonathan M. Weinberg. (2007) Analytic Epistemology and Experimental Philosophy. *Philosophy Compass* 2:56-80.
- Alicke, Mark D. (2000) Culpable Control and the Psychology of Blame. *Psychological Bulletin* 126:556-74.
- Blair, Robert James Richard, Lawrence Jones, Fiona Clark, and Margaret Smith. (1997) The Psychopathic Individual: A Lack of Responsiveness to Distress Cues? *Psychophysiology* 34:192-98.
- Campbell, C. A. (1951) Is 'Free will' a Pseudo-Problem? *Mind* 60:441-65.
- Cover, J. A., and John O'Leary-Hawthorne. (1996) Free Agency and Materialism. In *Faith, Freedom, and Rationality*, edited by Jeff Jordan and Daniel Howard-Snyder, pp. 47-71. London: Rowman and Littlefield.
- Dennett, Daniel. (1984) I Could Not Have Done Otherwise - So What? *The Journal of Philosophy* 81:553-65.
- Doris, John M., Joshua Knobe, and Robert L. Woolfolk. (2007) Variantism About Responsibility. *Philosophical Perspectives* 21.
- Doris, John M., and Stephen Stich. (2005) As a Matter of Fact: Empirical Perspectives on Ethics. In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith. Oxford: Oxford University Press.
- Double, Richard. (1991) *The Non-Reality of Free Will*. New York: Oxford University Press.
- . (1996) *Metaphilosophy and Free Will*. New York: Oxford University Press.
- Ekstrom, Laura. (2002) Libertarianism and Frankfurt-Style Cases. In *The Oxford Handbook of Free Will*, edited by Robert Kane, pp. 309-22. New York: Oxford University Press.
- Feltz, A. and E. T. Cokely (2008). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* In Press, Corrected Proof.
- Feltz, Adam, Edward T. Cokely, and Thomas Nadelhoffer. (2009) Natural Compatibilism Versus Natural Incompatibilism: Back to the Drawing Board. *Mind and Language* 24:1-23.

- Feinberg, Joel. (1970) *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton: Princeton University Press.
- Fischer, John Martin. (1994) *The Metaphysics of Free Will: An Essay on Control*. Aristotelian Society Monographs: Blackwell Publishers.
- . (2006) *My Way: Essays on Moral Responsibility*. Oxford: Oxford University Press.
- . (2007) Compatibilism. In *Four Views on Free Will*, pp. 44-85. Oxford: Blackwell Publishing.
- Fischer, John Martin, and S.J. Ravizza, Mark (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge Studies in Philosophy and Law: Cambridge University Press.
- Frankfurt, Harry G. (1969) Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy* 66:829-39.
- . (1988) *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press.
- Goldman, Alvin, and Joel Pust. (1998) Philosophical Theory and Intuitional Evidence. In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by Michael R. DePaul and William Ramsey, pp. 179-201. Oxford: Rowman and Littlefield Publishers, Inc.
- Gosling, S. D., P. J. Rentfrow, et al. (2003). A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality* 37(6): 504-528.
- Greene, Joshua. (2002) The Terrible, Horrible, No Good, Very Bad Truth About Morality and What to Do About It. In *Philosophy*: Princeton University.
- . (2003) From Neural 'Is' to Moral 'Ought': What Are the Moral Implications of Neuroscientific Moral Psychology. *Neuroscience* 4:847-50.
- Gutting, Gary. (1998) Historical and Metaphilosophical Introduction. In *Rethinking Intuitions*, edited by Michael R. DePaul and William Ramsey, pp. 3-17. Oxford: Rowman & Littlefield Publishers.
- Haidt, Jonathan. (2001) The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108:814-34.
- Honderich, Ted. (1993) *How Free Are You? The Determinism Problem*. 2nd ed. Oxford: Oxford University Press.
- Inwagen, Peter van. (1983) *An Essay on Free Will*. Oxford: Oxford University Press.

- Jackson, Frank. (1998) *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.
- Kane, Robert. (1996) *The Significance of Free Will*. New York, Oxford: Oxford University Press.
- . (1999) Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *Journal of Philosophy* 96:217-40.
- . (2005) *A Contemporary Introduction to Free Will*. New York: Oxford University Press.
- . (2007) Libertarianism. In *Four Views on Free Will*, pp. 5-44. Oxford: Blackwell Publishing.
- Kauppinen, Antti. (2007) The Rise and Fall of Experimental Philosophy. *Philosophical Explorations* 10:95-118.
- Knobe, Joshua. (2003) Intentional Action and Side Effects in Ordinary Language. *Analysis* 63:190-94.
- . (2006) Experimental Philosophy. *Philosophy Compass* 2:81-92.
- Knobe, Joshua, and John M. Doris. (2010) Responsibility. In *The Handbook of Moral Psychology*, edited by John M. Doris et al. Oxford: Oxford University Press.
- Knobe, Joshua, and Shaun Nichols. (2008) An Experimental Philosophy Manifesto. In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, p. 256. Oxford: Oxford University Press.
- Kunda, Ziva. (1990) The Case for Motivated Reasoning. *Psychological Bulletin* 108:480-98.
- Lerner, Jennifer S., Julie H. Goldberg, and Philip E. Tetlock. (1998) Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin* 24.
- Lycan, William. (2003) Free Will and the Burden of Proof. In *Proceedings of the Royal Institute of Philosophy for 2001-02*, edited by Anthony O'Hear, pp. 107-22. Cambridge: Cambridge University Press.
- Nadelhoffer, Thomas, and Eddy Nahmias. (2007) The Past and Future of Experimental Philosophy. *Philosophical Explorations* 10:123-49.
- Nagel, Thomas. (1979) Moral Luck. In *Mortal Questions*, edited by Thomas Nagel, pp. 24-38. Cambridge: Cambridge University Press.

- Nahmias, Eddy (2006) Folk Fears About Freedom and Responsibility: Determinism Vs. Reductionism. *Journal of Cognition and Culture* 6:215-37.
- Nahmias, Eddy, Justin D. Coates, and Trevor Kvaran. (2007) Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions *Midwest Studies in Philosophy* 31:214-42(29).
- Nahmias, Eddy, Stephen G. Morris, Thomas Nadelhoffer, and Jason Turner. (2005) Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility. *Philosophical Psychology* 18:561-84.
- . (2006) Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research* 73:28-53.
- Nahmias, Eddy, and Dylan Murray. (2010) Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions. In *New Waves in Philosophy of Action*, edited by Jesus Aquilar, Andrei Buckareff and Keith Frankish, pp. 189-217. Hampshire: Palgrave Macmillan. (The manuscript quoted from is available at <http://www2.gsu.edu/~phlean/papers.html>).
- Nelkin, Dana K. (2007) Do We Have a Coherent Set of Intuitions About Moral Responsibility? *Midwest Studies in Philosophy* 31:243-59.
- Nichols, Shaun. (2004) The Folk Psychology of Free Will: Fits and Starts. *Mind and Language* 19.
- . (2004) *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
- Nichols, Shaun, and Joshua Knobe. (2007) Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous* 41:663-85.
- Nichols, Shaun, Stephen Stich, and Jonathan M. Weinberg. (2003) Metaskepticism: Meditations in Ethno-Epistemology. In *The Skeptics*, edited by S. Luper, pp. 227-47. Aldershot, England: Ashgate Publishing.
- O'Connor, Timothy. (1995) *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press.
- . (2002) *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- Pereboom, Derk. (2001) *Living without Free Will*. Cambridge: Cambridge University Press.
- . (2007) Hard Incompatibilism. In *Four Views on Free Will*, pp. 85-126. Oxford: Blackwell Publishing.

- Pink, Thomas. (2004) *Free Will: A Very Short Introduction*. New York: Oxford University Press.
- Prinz, Jessie. (2007) *The Emotional Construction of Morals*. Oxford: Oxford University Press.
- Railton, Peter Albert. (2003) Morality, Ideology, and Reflection: Or, the Duck Sits Yet. In *Facts, Values, and Norms: Essays Toward a Morality of Consequence*, edited by Peter Albert Railton. Cambridge: Cambridge University Press.
- Ramsey, William. (1998) Prototypes and Conceptual Analysis. In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by Michael R. DePaul and William Ramsey, pp. 161-79. Oxford: Oxford University Press.
- Rawls, John. (1971) *A Theory of Justice*. Oxford: Oxford University Press.
- Sarkissian, Hagop, Amita Chatterjee, Felipe De Brigard, Joshua Knobe, Shaun Nichols, and Smita Sirker. (2010) Is Belief in Free Will a Cultural Universal? *Mind and Language* 25:346-58.
- Smart, D., and G. Loewenstein. (Forthcoming) The Devil You Know: The Effects of Identifiability on Punitiveness. *The Journal of Behavioural Decision Making*.
- Smilansky, Saul. (2000) *Free Will and Illusion*. Oxford: Oxford University Press.
- . (2003) Compatibilism: The Argument from Shallowness. *Philosophical Studies* 115(3):257-82.
- Sommers, Tamler. (2010) Experimental Philosophy and Free Will. *Philosophy Compass* 5:199-212.
- Sosa, Ernest. (2009) A Defense of the Use of Intuitions in Philosophy. In *Stich and His Critics*, edited by Dominic Murphy and Michael Bishop: Blackwell Publishers.
- Stanovich, Keith E. (2004) *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: University of Chicago Press.
- Stich, Stephen. (2009) Replies. In *Stich and His Critics*, edited by Dominic Murphy and Michael Bishop, pp. 190-253. Oxford: Blackwell Publishing.
- Stich, Stephen, and Jonathan M. Weinberg. (2001) Jackson's Empirical Assumptions. *Philosophy and Phenomenological Research* LXII.
- Strawson, Galen. (1986) *Freedom and Belief*. Oxford: Oxford University Press.

- Strawson, Peter. (1962) Freedom and Resentment. *Proceedings of the British Academy* XLVIII.
- Turner, J. and E. Nahmias (2006). Are the Folk Agent-Causationists? *Mind and Language* 21(5): 597-609.
- Vargas, Manuel. (2004) Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly* 85:218-41.
- . (2005) The Revisionist's Guide to Responsibility. *Philosophical Studies* 126:399-429.
- . (2006) Philosophy and the Folk: On Some Implications of Experimental Work for Philosophical Debates on Free Will. *Journal of Cognition and Culture* 6:239-53.
- . (2007) Revisionism. In *Four Views on Free Will*. Oxford: Blackwell Publishing.
- Viney, Wayne, David A. Waldman, and Jacqueline Barchilon. (1982) Attitudes toward Punishment in Relation to Beliefs in Free Will and Determinism. *Human Relations* 35:939-49.
- Weinberg, Jonathan M. (2007) How to Challenge Intuitions Empirically without Risking Skepticism. *Midwest Studies in Philosophy* XXXI:318-43.
- Weinberg, Jonathan M., Shaun Nichols, and Stephen Stich. (2001) Normativity and Epistemic Intuitions. *Philosophical Topics* 29:429-60.
- Wolf, Susan. (1980) Asymmetrical Freedom. *Journal of Philosophy* 77:157-66.
- . (1990) *Freedom within Reason*. Oxford: Oxford University Press.
- Woolfolk, Robert L., John M. Doris, and John M. Darley. (2006) Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition* 100(2):283-301.