Deep Selves in Moral Responsibility

Challenging the Realist Assumption

By: **Mischa Durham Davenport**

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Philosophy

University of Sydney

Faculty of Arts and Social Sciences

Department of Philosophy

2019

Abstract

Deep selves, in one way or another, feature significantly within our folk-psychology and play a particularly central role in our moral responsibility practices. Our intuitions about others' deep selves explain an otherwise complex pattern of responsibility attribution, with agents being held responsible for behaviour that in some way reflects the contents of their deep selves and getting off the hook for behaviour that doesn't. Philosophers who have addressed the deep self concept directly have typically done so on the assumption that the object of these intuitions is a real thing - a natural psychological kind in the agents towards whom our deep self intuitions are directed. What I will put forward in this thesis is a challenge to this realist assumption and an alternative framing of the deep self as constituted by response-dependent properties. I begin by introducing the basic concepts: the deep self, attributability and a 'Strawsonian reversal' in the relationship between our reactive attitudes and the facts of moral responsibility. Chapter II is concerned with the principal philosophical accounts of the deep self and argues that none provides a viable account of response-independent properties capable of constituting the deep self as a natural kind. Chapter III is concerned with empirical investigations bearing both directly and indirectly on the deep self concept, proposing a cognitive account of deep self intuitions as products of our folk-psychology and causal reasoning. Chapter IV examines some of the practical implications of this model and presents arguments in favour of abandoning the realist assumption.

Acknowledgements

I would like to thank my primary supervisor, Assoc. Prof Luke Russell, for his invaluable comments, advice and support in the preparation of this thesis.

I would also like to thank my auxiliary supervisor, Prof Paul Griffiths, for his contributions to this thesis.

Finally, I would like to acknowledge the following people for their helpful comments on various parts of this thesis: Adam Piovarchy; Jesse Miller; Arnaud Pocheville; Kate Lynch; Pierrick Bourrat; Stefan Gawronski; Elena Walsh; Daniel Finlay; Karola Stotz; Isobel Ronai; Wesley Fang.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any other degree or other purposes.

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

Mischa Davenport

Table of Contents

Chapter I: Deep Selves and Moral Responsibility

§1) Introduction

When philosophers talk about 'deep selves' – whether they use that or any other terminology – they tend to do so on the assumption that what they are talking about is a real thing. It is, after all, such a deeply familiar concept that it's hard to think of it as anything else. Even in our folk theories, whether it be the 'deep', the 'true', the 'authentic' or the 'inner' self, the underlying idea is undeniably pervasive – not just in the new-age pop-psychology of discovering one's 'inner self', but embedded within our folk-psychology itself and our very concept of agency.[1] It is what we appeal to when we ask who somebody *really* is, what they *really* care about, whether they *really* meant to do something. A significant amount of our attention as folk psychologists seems to be directed towards distinguishing other agents' 'true' features from more superficial ones, constructing in the process models of 'deep selves' that inform our interactions with them. Moreover, our collective repository of folk-psychological wisdom is full of propositions concerning the deep self: generalisations about when the deep self may be revealed to us (such as 'in vino veritas', or the idea of 'true character' being revealed under pressure) or indeed about when it might not be (captured in the idea that an agent might be 'not himself' when drunk, or in a stressful situation).[2]

The concept is no less central to philosophical theories of moral responsibility, where some version of the deep self typically serves as a foundation for attributions of moral responsibility. Behaviour[3] is thought to be attributable to an agent (meaning that she can be subject to praise or blame on the basis of that behaviour) to the extent that it is caused by, or reflects or expresses, her deep self. The role that the deep self plays in such theories is twofold.

[1] For an overview, see (Sripada, 2017a).

[2] There are two importantly different meanings of the term 'folk-psychology' that need to be distinguished, drawing on a distinction pointed out by Stich and Nichols (2003). Folk-psychology can refer to both (i) the way in which we keep track of, explain and predict others' behaviour, understood as a set of cognitive capacities, and (ii) a kind of lay-psychological theory made up by platitudes about human behaviour. Whilst the deep self concept is present in both kinds of folk-psychology, the interesting possibility to be explored in this thesis is that the two do not properly converge. Indeed, even from the brief examples I have provided it should be clear that the stock of folk platitudes concerning the nature of the deep self cannot reflect the deep self that features in our mental models of others' behaviour for the simple reason that many of those platitudes are simply inconsistent with one another. This will be explored in more detail in Chapter II.

[3] The term 'behaviour' is used in this thesis as a catch-all term, intended to include a range of things for which an agent might plausibly be held responsible, including choices, acts, omissions, attitudes, etc. More will be said about the different categories that fall under the term 'behaviour' in Chapter II, but all that needs to be noted for now is that it extends beyond straightforward intentional action.

First of all, it fulfils the task, fundamental to any theory of moral responsibility, of establishing the necessary link between instances of praise- or blameworthy behaviour, which are necessarily in the past, and the agents towards whom our blame or praise are to be directed in the present. The deep self is a way of describing whatever underlying (blame- or praiseworthy) property or disposition of an agent is both responsible for some past behaviour and still enduring so as to justify our holding her accountable for it in the present. The idea that there is some underlying and continuing essence is what allows me to justify blaming your present self for a wrong done to me by your past self. In this sense the deep self concept is simply a way of picking out that which is continuous and constitutes an agent's identity for the purposes of moral responsibility.

The second role that the deep self concept fulfils relies not just on a general notion of continuous identity, but much more specifically on a distinction between 'deep' properties and more superficial ones. It serves to explain a wide range of responsibility-mitigating intuitions – instances where agents are *not* held responsible for their behaviour – on the basis that the behaviour in question is somehow *not* linked in the relevant way to an underlying ('*deep*') property or disposition. Of course not every case of responsibility mitigation involves recourse to the deep self concept. In many cases the more basic concept of intentional action is sufficient: accidental behaviour, for example, is often exempted from blame simply on the basis that the act in question is not intentional at all – perhaps the agent who bumped into me on the train simply fell, or was pushed. Where recourse to the deep self concept becomes necessary is in explaining how instances of *intentional* behaviour are nonetheless not attributable to an agent for the purposes of our moral responsibility judgments. Sometimes the responsibility mitigation in question is complete, as when we talk, for example, of 'compulsive' behaviour – behaviour that is in a basic sense intentional and yet that an agent performs despite herself. This is how some theorists would have us characterise addictive behaviour: actions undertaken in the pursuit of an addictive substance are clearly intentional, and yet there is a sense in which the agent is alienated from the intentions or desires in question so as to not really be responsible for them. Sometimes the responsibility mitigation in question is only partial, as when we speak of weak-willed behaviour: the agent who intentionally does some blameworthy act against her better judgment is perhaps less blameworthy than the agent who fully embraces the same blameworthy course of action. In each of these cases we are drawing a distinction between an agent's 'true' intentions – those which constitute her identity for the purpose of our moral

responsibility attributions, her 'deep self' – and those intentions in fact revealed through her actions.

The question to be answered in this thesis is the following: **What *is* the deep self, and what role should it play in our theory of moral responsibility?** In order to begin to sketch a response to this question, we must consider the philosophical context in which the deep self concept has been drawn on and the particular kind of theory of moral responsibility in which it plays a role. Specifically, recourse has been had to concepts like the deep self as one way of responding to the problem of determinism. In its simplest form, the thesis of determinism proposes that in our universe every state of affairs is determined by an antecedent state of affairs and the operation of the laws of nature. The truth of the thesis of determinism thus presents a problem for theories of moral responsibility that would rely on agents acting (exercising free will) so as to determine the future. There are two general positions that can be taken when faced with the possibility of determinism: (i) compatibilism, which claims that the truth of determinism is compatible with the existence of free will, and (ii) incompatibilism, which claims that the truth of determinism is incompatible with the existence of free will. The meaning of 'free will' is itself highly contested, but as the context in which it arises is the question of whether or not agents are morally responsible for their actions, we can understand 'free will' to simply refer to whatever minimal conditions are necessary for morally responsible agency.

Incompatibilists can be further separated into those who believe that some agents do sometimes exercise free will, and accordingly that the thesis of determinism is false (Libertarians) and those who believe that no agents ever exercise free will. The latter group may be broken up into those who believe the absence of free will to be a result of the truth of the thesis of determinism (Hard Determinists) or who believe that there is no free will even if the thesis of determinism turns out to be false (Hard Incompatibilists).[4] Libertarians, on the other hand, may rely on different models of the kind of control required for free will – arguing on the basis either of a 'Principle of Alternate Possibilities'[5] or of some version of sourcehood libertarianism – that some agents to in fact sometimes exercise the relevant kind of control over their actions to constitute morally responsible agency. It is not, however, with incompatibilist theories that we will be primarily concerned. Firstly, it is typically on the compatibilist side of the philosophical debate that the deep self concept arises in the first place. Additionally though,

[4] Generally this latter claim is based on arguments that a given kind of indeterminacy – e.g. quantum indeterminacy – is just as problematic for free will as determinism itself.

[5] See (Frankfurt, 1969).

as pointed out by Strawson (1962), the truth or falsity of the thesis of determinism is simply not the kind of fact that is capable of altering our moral responsibility intuitions (which are at the root of what he describes as the 'reactive attitudes'). The very existence of the folk concept of the deep self, as will become clear in Chapters II and III, demonstrates a folk theory of moral responsibility that is indifferent as to the potentially determined nature of our universe. This is further supported by empirical investigations into folk intuitions on the subject of compatibilism: even in a determined universe, what people are really interested in is whether the causal pathways leading to some action operate *through* an agent's deep self or bypass it, again highlighting the centrality of the deep self to moral responsibility.[6]

Turning, then, to the compatibilist side of the debate, we can identify two very different models of the kind of control required for free will morally responsible agency. The first kind of control – sometimes referred to as 'regulative control'[7] or the aforementioned 'Principle of Alternate Possibilities – requires for free agency that an agent be able to do otherwise than she in fact does. This runs into some fairly straightforward problems when faced with the thesis of determinism: the 'classical incompatibilist' argument is essentially that if determinism is true then no agent has access to the kinds of alternatives required for regulative control and thus free will and morally responsible agency. Some alternate possibilities compatibilists have attempted to respond to this argument with a so-called 'conditional analysis',[8] arguing that an agent has the relevant freedom to do otherwise if, given a different past in which she wanted to do otherwise, she would in fact have done otherwise. Whilst this may appear to give a somewhat satisfying account of the content of the relevant freedom, the conditional analysis has generally been abandoned for the reason that it sometimes gives surprising results. In many cases agents may be constrained to act in the way they do as a result of compulsive desires or aversions, and whilst it may be true to say that given a different past in which the agent had different desires she would have acted differently, this does not justify the conclusion that in the actual world the agent was free to act as she did.[9] The conditional analysis, it seems, may obscure the very factors that make certain actions unfree in the first place.

[6] See, for example, (Murray & Nahmias, 2014) or (Knobe & Nichols, 2017).

[7] See (Fischer & Ravizza, 1998)

[8] See, for example, (Hume, 1978, p. 73).

[9] See (van Inwagen, 1983, pp. 114–119)

A second kind of control, sometimes referred to as 'guidance control',[10] is the foundation of a second kind of compatibilism: sourcehood compatibilism. The central idea of sourcehood compatibilism is that an agent acts freely, and is morally responsible for her actions, when she is the source of those actions. A straightforward incompatibilist response is to point out that, if the thesis of determinism is true, no agent is the *ultimate* source of her actions, because any action can be traced back along causal pathways beginning outside the agent and merely operating through her. The challenge then for sourcehood compatibilists is to provide an account of *mediate* sourcehood – something less than ultimate sourcehood – that describes a way in which determined causal pathways may operate *through* an agent's psychology or mental states so as to produce something we would recognise as free, morally responsible agency. This is the work purportedly done by the deep self concept: identifying a certain (privileged) subset of an agent's psychology or mental states which are capable of producing morally responsible agency.[11]

Whilst the deep self has been characterised in a range of different ways, which will be explored in more depth in Chapter II, I wish to begin by introducing two different ways of conceiving the relationship between our moral concepts and the world: Natural Kind Realism (the idea that our moral concepts correspond to natural kinds in the world) and Anti-Realism (the idea that our moral concepts *don't* correspond to natural kinds in the world) about moral properties.[12] It is my contention that source compatibilists either explicitly or implicitly adopt

---

[10] See (Fischer & Ravizza, 1998)

[11] It should be noted that the folk concept of the deep self – that is, a particular set of intuitions about when behaviour is attributable to agents – is a general intuition that could effectively be grafted on to any philosophical account of moral responsibility, be it Libertarian or Compatibilist. Alternate possibility Libertarians, for example, might suggest that our deep self intuitions are simply tracking those cases in which a given action was truly undetermined, such that the action in question truly was a reflection of the agent's deep self. If indeed there is a deep self – some core of an agent's practical identity - then it would only make sense to suggest that it would be expressed in just those situations in which multiple alternatives are genuinely open to an agent.

[12] While we are primarily interested in this distinction in the context of sourcehood compatibilism, it has also been suggested that these alternatives present themselves equally to the Libertarian – or at least that Anti-Realism is not necessarily incompatible with Libertarianism. This argument, raised by (Todd, 2016), depends largely on what I consider to be inadequate formulations of the Anti-Realist position (a point that I will address in §4 of this Chapter). On an adequate formulation of the Anti-Realist position the anti-realist libertarian approach becomes either incoherent or simply implausible. Such an approach could be formulated in one of two ways. The first would involve claiming that the counterfactual possibility necessary for moral responsibility is not a matter of objective fact, but merely psychologically constructed, which would be inconsistent with the strict metaphysical construal of 'counterfactual possibility' required for the incompatibilist position. The second would involve the claim that, whilst the counterfactual possibility in question is a matter of metaphysical fact, it is simply a contingent fact of our moral practices that this is the feature we have taken to be constitutive of morally responsible agency, which would seem wildly implausible in light of the obviously central role such counterfactual possibility would necessarily play in any theory of agency. Just what an account of the nature of the deep self might look like within this framework is extremely unclear.

a realist assumption about the deep self. It is sourcehood compatibilism – and the need for a compatibilist criterion of morally responsible agency – that has driven the development of philosophical accounts of the deep self. Rejecting the principle of alternate possibilities, compatibilists typically argue that morally responsible agency consists in a particular kind of relationship between certain (privileged) elements of an agent's psychology and her behaviour: when behaviour is caused by, or on some accounts merely consistent with, the relevant subset of an agent's mental states, then that behaviour may rightly be attributed to the agent in question for the purposes of moral responsibility. Essentially the realist sourcehood compatibilist believes that our deep self concept identifies a natural psychological kind – a set of mental states identified by certain objective (response-independent) properties – such that any claim about the content of a particular agent's deep self has its objective truth conditions in facts about that agent's psychology. One consequence of deep self realism in general is a corresponding objectivism about moral responsibility facts – the implications of which I will return to consider in Chapter IV.[13]

The position, as yet undeveloped, that I will advance in this thesis represents an alternative to this traditional picture and the realist assumption on which it relies. According to the anti-realist account that I will propose, the features that constitute morally responsible agency – the apparent objects of our deep self intuitions – are fundamentally response-dependent. This would mean that there is no objective (response-independent) property common to all instances of morally responsible agency, and that our deep self intuitions, rather than tracking a natural psychological kind in the agents to whom they are directed, in fact represent the products of a cognitive mechanism that *constructs* the deep self as a feature merely of each individual's folk-psychological reality. The resulting criteria for morally responsible agency will necessarily be compatible with the truth of determinism because the truth of determinism would not alter the underlying psychological mechanisms that produce the deep

[13] Admittedly, the Realist/Anti-Realist distinction is more complex than the simple dichotomy presented here. Whilst the strict Realist would argue that there is a natural kind property that constitutes morally responsible agency because of its *independent* relationship to agency itself, and the strict Anti-Realist would argue that there is no such natural kind property but rather a response-dependent property, there exists a middle ground according to which there may be a real, natural kind property present in all instances of morally responsible agency, but which is only considered relevant to moral responsibility because our moral practices *happen* to track that particular property (with the implication that a different set of moral practices might just as easily have identified a *different*, but equally real, feature as relevant for moral responsibility). Whilst some deep self theorists may not explicitly state which position (Realist or middle-ground) they are adopting, they invariably claim that the privileged relationship between a certain (natural) kind of mental states and morally responsible agency is not merely due to the importance contingently attributed to such mental states within our moral practices, insisting instead that the mental states in question actually enjoy an independently privileged status given its special causal role within human agency.

self intuitions in question. What counts as morally responsible agency – those behaviours that can be attributed to an agent's deep self, and therefore the *contents* of any agent's deep self – will vary depending on certain features of the observer.

The more specific account that I will defend explains our deep self intuitions as the products of a very basic cognitive mechanism – the causal cognition at work in our folk-psychology – which postulates unobserved causal essences as features of more general causal models.[14] This is of course not the only explanation of our deep self intuitions open to the anti-realist. Other accounts might refer, for example, to the instrumental pay-off of making deep self attributions (and the associated blaming responses) in certain situations. On the account that I will propose, the reason that deep self properties are response-dependent is that the way in which these causal essences (deep selves) are attributed depends on factors specific to a particular observer. One consequence of this approach, in contrast to the realist approach outlined above, is a relativism about moral responsibility facts. To give a basic example of the account I am proposing, consider addictive behaviours as a contested category in terms of moral responsibility judgments: some may insist that agents suffering from an addiction are not responsible for their behaviour because it fails to reflect their deep selves; others, who don't accept the disease model of addiction, would instead see such behaviour as expressing some (morally blameworthy) element of the addicts' deep selves. A realist approach to this kind of conflict would be to claim that there is an objective fact of the matter as to whether an agent's deep self is reflected in the behaviour in question, and that one set of moral responsibility intuitions must be wrong. The anti-realist approach that I will advance would instead claim that there is no objective fact of the matter about the content of an addict's deep self, and seek to understand which relevant features of each observer's perspective are responsible for the difference in deep self intuitions. Whilst the argument I am primarily concerned to advance is *against* deep self realism, in Chapter IV I will consider and compare a number of possible anti-realist views including relativism, expressivism or error theory about deep self attributions.

The defence of this anti-realist approach will proceed as follows. In the remainder of this chapter, I will begin by introducing the concept of the deep self in greater depth, elaborating on the role that it plays within our folk psychology and outlining the general form that it takes in contemporary philosophical accounts (§2). I will then clarify the precise aspect of moral responsibility – *attributability* – in which the deep self concept plays such an important role

[14] In Chapter III I explain why I am inclined to prefer this particular explanation of our deep self intuitions.

(§3), before elaborating on the realist/anti-realist distinction through a discussion of the Strawsonian reversal (§4). Chapter II will be devoted to an examination of contemporary philosophical accounts of the deep self, highlighting the shared realist assumption on which a range of competing accounts rely and arguing that none proposes a viable response-independent criterion capable of constituting the deep self as a natural psychological kind. In Chapter III I turn to an emerging body of empirical literature on folk attributions of moral responsibility, including some specific investigations of the deep self concept, in order to support my anti-realist account of the cognitive processes that result in our deep self intuitions. Finally, Chapter IV will consider some of the practical implications of my proposed model of the deep self and present arguments in favour of abandoning the realist assumption in our deep self talk.

§2) Introducing the Deep Self Concept

Some possible initial scepticism notwithstanding (from those of us for whom the term 'deep self' is synonymous with a particular kind of pop-psychology) the basic idea of the deep self ought, by all accounts, to strike us as highly intuitive. The deep self concept, in some general form, manifests itself when we reflect on the kind of person that somebody *really* is; on what they *really* want; on whether somebody, deep down, is genuinely committed to something, or authentically cares about it. A deep self is what we implicitly imagine when we ask of another's behaviour whether it *really* reflects on some important aspect of their character or is instead in some way an outlier – unreflective of who they '*really'* are. We might imagine that a particularly strict parent, who shows no obvious outward signs of affection towards a child, as nonetheless really, deep down, loving that child – perhaps their affection even explains the behaviour in question, we might say, as it shows that they ultimately want what's best for their child. Similarly we might imagine a politician who is perpetually professing his commitment to the working class, endlessly being seen in hard-hats and hair-nets, and whom we nonetheless believe to be motivated, *deep down*, by nothing but self-interest. An investment banker might show all the outward signs of being happy, and yet we are capable of maintaining that, deep down, she must be deeply unsatisfied by the trappings of her wealth and success. The idea of an agent's observable behaviour in some way belying her true nature is so familiar that each of the examples just mentioned represents a kind of cultural trope.

We are constantly making judgments about who people *really* are – as distinct from who they appear to be – that demonstrate our folk belief in the existence of a deep self. Its most obvious function becomes clear when we reflect on moral responsibility: on the role of

underlying mental states, attitudes or dispositions that make an agent's behaviour attributable to her for the purposes of moral blame or praise (or indeed that make those actions, as expressions of underlying mental states, blameworthy or praiseworthy to begin with). Imagine an otherwise loving husband who uncharacteristically snaps at his wife: does this reveal a deep, underlying (morally blameworthy) element of his character, or was it simply a stressful situation in which an otherwise virtuous agent made a mistake? Which behaviour to we attribute to the agent's *deep* self – the self that we refer to for the purposes of moral evaluation? In assessing an agent's praise- or blameworthiness for some behaviour, we first need some way to *link* the behaviour in question (necessarily in the past) to the agent who is to be the target of our (present) moral assessment – and the deep self, as a set of stable, underlying mental states/attitudes/dispositions, *constitutes* this link.

What the deep self concept presupposes, of course, is a kind of superficial self – a set of properties or behaviours that are not properly attributable to an agent. But how to conceive of this disconnect between deep and superficial selves? In some cases the distinction may be relatively simple, with the deep self expressed in an agent's conscious decisions and her superficial self in behaviours that are accidental, uncharacteristically negligent or perhaps even coerced. In other cases though, the relationship between the deep and superficial self may be more complex: an agent may consciously endorse particular behaviours – consciously identify with certain of her motivations – and turn out to be mistaken or self-deceived; perhaps her deep self is not revealed but hidden by all of the conscious processes of thought. Of course there are simpler ways of conceptualising the deep self.

One appealingly simple model, which might be described as 'classical compatibilism', suggests that one's true self is constituted by one's strongest motivations, where 'strongest' simply means those motivations that move an agent to action.[15] On this kind of account *all* intentional actions would reflect an agent's deep self, and the only deep/superficial distinction would be between weaker and stronger motivations. The much simpler model that we are left with confines the superficial self to an agent's words – the values and attitudes she might claim to hold – as opposed to her deep self which is manifested in action. Insofar as there can be any uncertainty around the content of an agent's deep self, it is the uncertainty involved in

[15] An example of classical compatibilism is the account of moral freedom and responsibility put forward by Hobbes, according to which an agent acts freely when she is unimpeded in doing what she wants, where what she wants is not analysed any further than to identify it as that which she would, in the absence of external impediments, choose to do (Hobbes, 1997, p. 108). The sense in which this constitutes freedom of will has however been described as 'deflationary' (McKenna & Coates, 2018).

determining the motivations behind an agent's action. There is no question, however, of a distinction between *sources* or *kinds* of motivation.

The problem with the classical compatibilist account is that it does a poor job of accommodating many of our intuitions about moral responsibility, as we in fact regularly do differentiate between different motivations on grounds other than their respective strength, such that an agent's strongest motivation (that which is effective in action) is *not* always attributable to her. We may think a certain behaviour is not attributable to an agent's *deep* self because it is caused by motivations that, for example, are merely a product of her acculturation, or by motivations that run counter to and override her *real* values and attitudes, or because of any of a range of extenuating circumstances that might mean that her behaviour does not accurately reflect her deep self (instances of compulsive behaviour, behaviour that is compelled by others or by a situation, or behaviour performed under the influence of intoxicating substances). Addiction, as it is commonly understood, is one such example: an agent's behaviour can be explained in terms of a particular motivation (in this case, the desire for a certain drug) where we understand the motivation in question to be *external* to the agent and so not properly attributable to her.

Contemporary deep self accounts do a much better job of accommodating these intuitions. They do so by identifying a specific subset of an agent's mental states or attitudes (motivations) as those relevant for the purposes of moral responsibility. The defining feature of morally responsible agency is taken to be a particular relationship (options include an *expression* relationship, a *concordance* relationship, a *causal* relationship, etc.) between a certain subset of an agent's mental states and her behaviour. What those specific mental states are is the subject of some controversy, but a few competing accounts have characterised them as either cares, commitments, second-order desires or evaluative judgments. The moral responsibility-defining 'freedom', on this kind of realist, compatibilist account, is defined not in terms of counterfactual possibility but in terms of the absence of impediments to expression of an agent's deep mental states in behaviour. This fundamentally compatibilist concept of freedom rejects the necessity of alternate possibilities (the 'freedom to do otherwise') in favour of a *criterion of internal sourcehood*, according to which free acts are simply those that flow in the right way from particular elements of an agent's psychology.

Whilst deep self theorists agree on this general idea of a criterion of internal sourcehood, there is much debate between theorists as to how exactly the relevant subset is to be characterised. Though proposals vary widely, they can generally be described as falling within

one of two dominant strands: a 'desires-based' strand that has its recent origins in the works of Harry Frankfurt, and a competing 'reasons-based' or evaluative strand whose origins can be found in the works of Gary Watson. Each represents an attempt to identify a subset of psychological states that are the focus of our moral responsibility intuitions, in order to accommodate the intuitions mentioned above involving non-attributable intentional actions.

The Frankfurtian approach begins, for our purposes, with 'Freedom of the Will and the Concept of a Person',[16] though it undergoes significant modifications both in Frankfurt's own later contributions[17] and in more recent contributions including those by Shoemaker (2003), Arpaly and Schroeder (2013) and Sripada (2016). What such accounts have in common is the characterisation of the psychological elements relevant for moral responsibility as a particular set of *desires*, described in later accounts as *cares* (a kind of 'foundational' desire providing intrinsic as opposed to instrumental motivation). Whilst on compatibilist accounts of moral responsibility all behaviour may be thought of as *determined*, deep self accounts attempt to distinguish between causal chains operating *through* an agent's deep self from those that operate in opposition to it, overwhelming or negating the agent's underlying dispositions. Actions produced in the former way, though still the products of determined causal mechanisms, may be thought of as *authenticated* by the agent's deep self and therefore attributable to her for the purposes of moral responsibility.

The Watsonian approach, on the other hand, begins as a response to Frankfurt's account in 'Free Agency',[18] though Watson later acknowledges certain shortcomings in his account.[19] Later contributions that can be seen as refining or modifying the Watsonian approach include those by Scanlon (1998) and Smith (2005).[20] What they have in common is the characterisation of the relevant psychological elements as a form of evaluative judgment, later more specifically described as evaluative *commitments*: 'foundational' evaluative beliefs that provides agents with a special kind of (internal, intrinsic) reason for action. Actions expressive of these commitments can be thought of as *authorised* by the agent's deep self, and therefore attributable to those agents.

[16] (Frankfurt, 1971)

[17] (Frankfurt, 1982, 1987, 1992)

[18] (Watson, 1975).

[19] (Watson, 1987a)

[20] See also (Wolf, 1990), (Nelkin, 2011).

Given the different ways in which each approach has characterised the deep self – the different mental states that each would describe as constituting the deep self – one might expect the two competing approaches to produce different deep self attributions, and in some instances this is true. In the case, for example, of an agent torn between the evaluative belief that pre-marital sex is wrong and the strong desire to have pre-marital sex, proponents of either a Watsonian or a Frankfurtian account might plausibly give different answers to the question of which mental state/attitude reflects the agent's deep self. There are many other cases though where they produce exactly the same deep self attributions. Both, for example, take the 'unwilling addict' as a core example of non-attributable intentions and as a source of evidence for their respective deep self characterisations, as the addict's resistance to or alienation from the motivations produced by his addiction can be cashed out either in terms of his 'true' evaluative judgments or indeed his 'true' desires.

The argument that I will present in Chapter II is that deep self accounts in both strands face a kind of trade-off. A more precise psychological description of the kind of mental state that constitutes the deep self is on the one hand more likely to attach to a natural psychological kind, but at the same time it results in a theory of the deep self that is less flexible and more likely to lead to counter-intuitive results. As a consequence, though such accounts may be able to argue that the deep self that they describe is a natural psychological kind, they will also have to accept that the deep self in question is not the same as the object of our deep self intuitions. A less precise psychological description of the kind of mental state in question produces a more flexible theory of the deep self, able to accommodate the full range of deep self intuitions, but the vague mental states it refers to are correspondingly less likely to identify any natural psychological kind.

§3) The Moral Responsibility Concept at Stake: Attributability

Having provided a brief outline of the first central concept in this thesis – that of the deep self – the next step is to provide a satisfactory picture of the aspect of moral responsibility in which it plays a role. Our blaming and praising responses are determined by many other factors than simply our deep self intuitions, and so it is worth establishing how these various factors fit together. The first important distinction to make is between our *blaming* intuitions and our *punishing* intuitions, recognising that the two do not always go hand in hand. As pointed out by Cushman (2008) our blaming intuitions are tied to judgments about the mental states of the agent in question, whereas our punishing intuitions are much more directly tied to causal

responsibility. Essentially, we tend to blame other agents to the extent that we think some harmful outcome is *desired* by them, but we tend to punish other agents according to their perceived causal responsibility for some harm. Moral responsibility, as I will be discussing it, will focus on blaming intuitions rather than punishing intuitions, as the latter may have little to do with judgments about an agent's mental states.

Our blaming intuitions themselves combine a number of distinct judgments. The judgment that I have discussed so far concerns the possibility of linking particular behaviours to agents via the attribution of deep mental states. A full picture of our moral responsibility intuitions (our blaming intuitions) would include the criteria according to which we make moral valence judgments – judgments concerning the moral valence of particular behaviours, or indeed the moral valence of particular deep mental states. Our blaming intuitions also rely on a judgment concerning the appropriateness of *holding* other agents responsible for behaviours or mental states that we might attribute to them – a judgment involving considerations of something like 'standing' to blame.[21]

For the purposes of this thesis however, I will be restricting my treatment of moral responsibility to the narrower question of *attributability*: on what basis are some behaviours attributed to an agent as reflecting her deep self, while others are not? Attributability judgments are, as just alluded to, only a single aspect of our broader moral responsibility practices.[22] In order to better delineate this aspect of our moral responsibility judgments from other important aspects of moral responsibility, as well as to clear up certain terminological ambiguities that might otherwise arise, it is worth considering a few prominent accounts of 'attributability' as a distinct concept.

---

[21] It may be objected that where something like 'standing conditions' are violated an agent does not cease to be accountability-responsible, but rather it is simply inappropriate for particular others to hold them accountable. This kind of objection relies however on a concept of accountability that is independent of an agent's accountability *to particular persons*. How many others need to lack standing in order to say that a particular agent goes from being accountability-responsible for some behaviour (even though it would be inappropriate for *some* other agents to hold her to account) to ceasing to be accountable for it altogether?

[22] An understanding of attributability may provide us with an understanding of the basis on which an attitude or action can be attributed to an agent as an element of her deep self, but it certainly would not account for why that element constitutes either a (morally praiseworthy) excellence or a (morally blameworthy) flaw – but the answers to these questions can effectively be separated from the more fundamental question of attributability for the purposes of this analysis (I say more fundamental because, in the order of our responsibility practices, *attributing* responsibility to an agent, on the basis of some kind of underlying mental state, is necessarily *prior* to judgments about the appropriateness of *holding* such an agent responsible for such mental states, and also, theoretically at least, distinct from the factors that ought to determine the moral *valence* of the attributed mental state – whether it be, on a realist account, moral facts concerning, e.g., the regard due to some individual or thing, or, on an anti-realist account, the requirements of the interpersonal demand, as defined within a context of moral practices).

Our starting point is Watson's 'Two Faces of Responsibility' (1996). In it Watson presents a distinction between *attributability* and *accountability* that is useful for clarifying the role of the deep self concept in moral responsibility judgments. The distinction is presented as a response to a criticism of the 'deep self model' of moral responsibility made by Wolf (1987).[23] Wolf's objection is that the posited relationship between underlying (deep) mental states and behaviour is on its own insufficient to explain our responsibility intuitions, noting that our responsibility judgments typically require not only that a criterion of 'internal control' be met (meaning that 'deep' mental states are expressed in behaviour) but also a criterion of external control – what Wolf describes as 'normative competence'.

The criterion of 'internal control' is what allows us to distinguish morally responsible behaviour from cases of, for example, compulsive behaviours – where an agent's deep mental states fail to express themselves.[24] The criterion of 'external control' on the other hand might explain how we treat agents like psychopaths. It is not the case, when we see a psychopath behave in some characteristically *evil* way, that we consider this behaviour to be contrary to her deep motivations – we do not imagine that he is behaving in this way 'despite himself'. Instead, we are likely to explain this behaviour by reference to a kind of deep, perverse motivation – that we may indeed consider a defining feature of psychopathy. The impairment of the psychopath is a failure in 'normative competence': a failure to perceive moral facts (however these might be cashed out) as being *reason-giving*. It is not because we take children, psychopaths and the insane to not be genuinely expressing their underlying motivations that we feel inclined to deny their responsibility, but because we view them as lacking the requisite normative competence in a way that makes holding them accountable inappropriate.[25]

[23] Wolf is in fact responsible for first describing the approaches of Frankfurt and Watson as 'deep self' theories of moral responsibility, though she also later refers to them as 'real self' accounts – see (Wolf, 1990).

[24] Perhaps at this stage it is worth pre-empting a concern of the reader about what might be described as 'blameworthy weakness of will', where we appear to hold agents responsible for behaviour even when the criterion of internal control appears not to be met. This particular case will be dealt with in some detail in Chapter II, where I will argue that weak-willed behaviour, rather than representing a failure of an agent's deep mental states to express themselves in action, is in fact a case in which the behaviour in question *is* deemed attributable to the agent *on the basis of a deep mental state* – where that mental state is not necessarily the problematic motivation in question, but an *insufficient* contrary motivation that *ought* to have defeated the problematic motivation in the first place.

[25] How exactly the 'inappropriateness' of such moral demands on agents like psychopaths or children is ultimately cashed out will depend on the theory of blame in question. On a communicative account of blame, it might be inappropriate to hold such agents responsible because they are simply incapable of understanding what is being communicated. On an account of blame as addressing an impediment to a particular interpersonal relationship, blame may be inappropriate because the relationship in question was never properly able to exist in the first place.

In response to these observations, Watson draws the aforementioned distinction between *attributability* (which he describes as the 'aretaic face' of responsibility, for which only the internal criterion of self-disclosure must be met) and *accountability* (which involves judging other agents to be accountable *to us* for their behaviour, based on a set of moral obligations to which we are mutually subject). Whilst attributability refers to a relationship between an agent's deep self and her behaviour, accountability refers to a kind of relationship between agents. It is thus possible for an agent to be an appropriate object of our attributability judgments without being an appropriate subject of moral demand or accountability judgments – a situation which Watson claims explains, amongst other things, our conflicted responses towards agents like psychopaths.

I make a brief parenthesis here to address the concept of 'answerability', which has been proposed by a number of philosophers[26] as an additional aspect of moral responsibility. Answerability for a particular behaviour means that an agent can reasonably be expected to defend or justify the underlying evaluative judgments that the behaviour in question reveals. The reason that I raise answerability only as a parenthesis is that the place of answerability within a broader theory of moral responsibility depends in a rather fundamental way on the way in which attributability is cashed out – in other words on how the deep self is characterised. Smith, for example, adopts a rationalist criterion for deep self attributability - meaning that an agent's deep self is constituted by evaluative judgments, and attributable behaviour is behaviour that in some way reflects these judgments. A natural consequence of this is that 'answerability' and 'attributability' are essentially coextensive: evaluative judgments are the kind of thing that one can be asked to defend or justify, and so any behaviour that is attributable to an agent (meaning that it is expressive of some evaluative judgment) is also behaviour for which she is answerable. For Shoemaker on the other hand, whose account of the deep self includes non-rational emotional commitments, or 'cares', there exists a whole range of behaviour which is attributable to agents (in the sense that it reflects those underlying non-rational commitments) but for which they are not answerable (they cannot reasonably be asked to defend or justify them). It is for this reason that the debate concerning the place of answerability within the broader field of moral responsibility need not concern us here. It will be most useful to retain the original distinction between attributability and accountability proposed by Watson, interpreted in such a way as to require no prior commitments concerning the nature of the deep

[26] See in particular (Scanlon, 1998), (Smith, 2007, 2008, 2012) and (Shoemaker, 2011, 2015b).

self. One might think of this as the 'content-neutral' formulation of the different faces of moral responsibility: functional descriptions of two different aspects of our moral responsibility judgments that do not depend on any prior assumptions about the nature of the deep self.

To reiterate then: *attributability* refers to a judgment concerning an agent's *intra*personal psychology, concerning the degree to which an agent's behaviour may be seen to express a particular kind of underlying mental state (where the question of exactly which mental states constitute an agent's deep self is left open); *accountability* refers to a judgment concerning an *inter*personal relationship of moral demands, concerning whether or not such demands (*of* one agent, or group of agents, *on* another) are appropriate (again, remaining neutral as to the question of what factors may be liable to render such demands inappropriate, be it the 'normative competence' of the agent or the 'standing' of those who would blame her). Attributability in this sense represents a kind of prerequisite to other aspects of our moral responsibility practices, and may be thought of as conceptually prior to questions of valence (judgments concerning the blame- or praiseworthiness of any attitude attributed to an agent's deep self) or of accountability. The role of the deep self concept within our moral responsibility practices more generally is as the proper object of our attributability judgments.

§4) From Realism to Anti-Realism: The Strawsonian Reversal

In order to clarify the contrast that I have in mind between the realist and anti-realist positions, I will begin by presenting the general form of what is sometimes called a 'Strawsonian' reversal. The reversal that the name refers to is in the 'order of explanation' between our reactive attitudes and the facts of responsibility. In its simplest form, the Strawsonian reversal involves a rejection of the standard realist understanding of moral responsibility according to which our reactive attitudes (such as resentment and gratitude) are explained or justified by reference to independent, objective moral responsibility 'facts'. On a realist account, our reactive attitudes may be thought of as playing an 'evidential' role with respect to moral responsibility – that is, they are in some way *tracking* a set of independent facts of moral responsibility. In order to understand just what this traditional realist picture is to be replaced with though, we will need to engage in some detail both with Strawson's own account and with its reception and interpretation within moral philosophy. I will endeavour in the first instance to reconstruct Strawson's argument concerning the nature of moral responsibility and its relationship to our reactive attitudes, before turning to consider its reception by successive moral philosophers. Having established and defended the version of

Strawsonian anti-realism that I will adopt with respect to moral responsibility in general, I will then conclude by considering exactly how this applies to the more specific concept of the deep self.

Strawson's approach to moral responsibility, as presented in 'Freedom and Resentment' (1962), is framed as an attempt to reconcile compatibilist and incompatibilists, but is perhaps more aptly described as a defence of the compatibilist thesis, making it in a sense similar to the works of Frankfurt (1971) and Watson (1975). The novelty of his approach is in turning away from abstract theoretical accounts of moral responsibility, both compatibilist and incompatibilist, in favour of a moral responsibility concept that he claims can be retrieved entirely from "the facts as we know them". The truth or falsity of the thesis of determinism, as an abstract metaphysical claim about a certain kind of freedom, is simply not part of the ordinary facts as we know them that provide us with our concept of moral responsibility. That concept is determined not by any set of independent moral facts (on which the truth of determinism may be claimed to have some bearing) but by a specific context of moral practices and by a set of intuitions which underpin what he describes as the 'reactive attitudes'. Strawson, on the optimist's behalf, presents two different concepts of freedom, one negative and one positive, arguing that the defining feature of morally responsible agency is not the positive freedom (actual freedom to do otherwise) that incompatibilist accounts make it out to be, but rather a negative freedom defined as "nothing but the absence of certain conditions the presence of which would make moral condemnation or punishment inappropriate" .

In order to recover this ordinary concept of moral responsibility, Strawson's first move is to bring the debate back into the realm of interpersonal relationships. The problem with both sides of the compatibilism debate, he suggests, is the tendency to argue from the detached perspective (from what he later labels the 'objective stance') and the solution is to immerse oneself in the psychology of non-detached, *participant* reactive attitudes: 'such things as gratitude, resentment, forgiveness, love and hurt feelings'. These, for Strawson, constitute the facts as we know them. Reactive attitudes, and the contexts in which we experience them – as responses to the perceived quality of will of those around us – are not to be taken as 'evidence' of independent facts about moral responsibility, but as the basic facts that are themselves *constitutive* of moral responsibility. Instances of morally responsible behaviour just are those behaviours that elicit reactive attitudes.

From these reactive attitudes Strawson proposes to establish the contours of the concept of moral responsibility: what situations are capable of eliciting the reactive attitudes, and what

considerations are capable of mitigating them? The latter considerations – responsibility-mitigating factors – Strawson divides into two kinds: those that may excuse or justify a particular instance of behaviour ('it was an accident', 'he didn't mean it', 'she had no choice,') as not truly expressive of a poor quality of will, and those that may lead us to exempt an agent, temporarily or permanently, from the whole range of our reactive attitudes ('he was drunk, 'she's not herself', 'he's not right in the head') despite the ill will otherwise manifest in their behaviour. Given this taxonomy of features capable of mitigating our reactive attitudes, Strawson concludes that the truth of the thesis of determinism is simply not the right *kind* of consideration to make us abandon our reactive attitudes.

The reversal that has taken place here is hard to pin down but extremely significant: whereas a standard realist approach to moral responsibility assumes our moral responsibility practices (our reactive attitudes etc.) to be grounded in independent facts about moral responsibility, Strawson's alternative proposal is that the facts of moral responsibility are grounded in our moral practices. Nothing can alter the facts about moral responsibility that does not involve altering our moral practices. The question of compatibilism is thus answered not by whether or not the truth of determinism would alter any abstract facts about moral responsibility, but by whether or not it would alter our concrete moral practices – whether it would cause us to stop experiencing the reactive attitudes.

Just how are we to interpret the new relationship between reactive attitudes and moral responsibility that Strawson seems to be promoting? One relatively straightforward way in which this reversal could be understood, in contrast to either a consequentialist (compatibilist) or Libertarian (incompatibilist) formulation of the realist position, is as follows:

> What these otherwise very different views [of the consequentialist and the libertarian] share is the assumption that our reactive attitudes commit us to the truth of some independently apprehensible proposition which gives the content of the belief in responsibility; and so either the search is on for the formulation of this proposition, or we must rest content with an intuition of its content. For the social-regulation theorist, this is a proposition about the standard effects of having and expressing reactive attitudes. For the libertarian, it is a proposition concerning metaphysical freedom. Since the truth of the former is inconsistent with the thesis of determinism, the consequentialist is a compatibilist; since the truth of the latter is shown or seen not to be, the libertarian is an incompatibilist.
>
> In Strawson's view, there is no such independent notion of moral responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they are responsible; <u>rather, they are responsible because we hold them responsible</u>. (Todd, 2016) citing (Watson, 2004a) except for the last line (underlined).

In practice however very few philosophers have interpreted the Strawsonian reversal in this way, by embracing the implication that the fact of another agent's moral responsibility is determined by our practice of holding her responsible. Watson, for example, in the original version of the passage quoted above, proposes a kind of expressivist interpretation, claiming that:

> [T]he idea (our idea) that we are responsible is to be understood by the practice, which is not itself a matter of holding some propositions to be true, but of expressing our concerns and demands about the treatments of one another" (Watson, 1987b)

One potential consequence of such an expressivist account of our moral intuitions and reactive attitudes is that any resulting claims about moral responsibility are simply not the kind of claim that is susceptible of being true or false. Some ambiguity is introduced by further formulations that involve a constitutive relationship:

> Strawson's radical claim is that "reactive attitudes" […] are **constitutive** of moral responsibility; to regard one another as responsible just is the proneness to react to them in these kinds of ways under certain conditions. There is no more basic belief which provides the justification or rationale for these actions. The practice does not rest on any theory at all, but rather on certain needs and aversions that are basic to our conception of being human. (Watson, 1987b, emphasis mine)

Let us begin then with an investigation of the 'constitutive' relationship in question. The most 'natural' or straightforward meaning of such a formulation, according to Todd, is as follows: "*that the fact that we have these attitudes towards one another makes us morally responsible – that is, makes us deserving of such attitudes, or 'appropriate targets' of such attitudes*" (Todd, 2016). The 'makes' in this formulation is on its face ambiguous as between a causal relationship and a somewhat vaguer sense of 'making it the case that…'. The difference can be seen through an analogy with a quality like redness: "*If you say that something's being red is constituted by its disposition to cause certain sensations in certain subjects, then you're saying, I think, that the fact that something causes those sensations makes it red*" (Todd, 2016). What is meant in this case is that a claim about the redness of some object amounts to a claim that that it does in fact have the relevant disposition to cause certain sensations in a certain kind of subject – the claim has its truth conditions in a set of facts concerning the sensations that subjects are disposed to experience. We may be independently interested to know what it is that *causes* an object to be red, which is not simply a feature of the object (perhaps its reflective qualities) but rather a combination of factors including the sensory apparatus of the observer and the ambient light conditions. These features can be thought of as causally responsible for

the object's redness because manipulating any one of them would make the object cease to produce the relevant sensation, therefore ceasing to be red.

Returning to claims about moral responsibility, the idea that moral responsibility is constituted by our reactive attitudes is perhaps best interpreted as follows: any claim about moral responsibility essentially amounts to a claim about reactive attitudes. Accordingly, if I claim that *S* is morally responsible for *a*, all I am really claiming is that 'one' (defined according to the context in which the claim is made) is disposed to experience certain reactive attitudes towards *S* on the basis of *a*. That is, the truth conditions for such a claim relate only to the disposition of *a* to cause a certain kind of reactive attitude. There is no question raised on this formulation of the Strawsonian reversal of the 'appropriateness' or 'fairness' of such reactive attitudes, and no reference to any objective, external reality of moral responsibility on which these reactive attitudes supervene. Of course it is not the case that our experiencing certain reactive attitudes *causes* an agent to be responsible. What causes any particular instance of moral responsibility, on the version of the Strawsonian reversal that I am proposing, is a combination of facts about the agent's behaviour, facts about the 'sensory apparatus' (the folk-psychological cognitive apparatus) of the observer and certain contextual facts including salient moral norms and social narratives. Intervening on any one of these features ought accordingly to be capable of making it the case that an agent either is or isn't morally responsible for some behaviour.

This particular formulation of the Strawsonian reversal, according to which moral responsibility is an entirely response-dependent property, has not received a great deal of support. Even Watson, who provides the 'constitutive' formulation referred to above, considers the theory to be 'incomplete in important respects', raising the possibility that what is needed in order to complete the theory may well undermine it. What Watson and others, including Tognazzini[27] and Wallace[28], propose to fill this apparent gap in the Strawsonian theory is an account of the 'propriety conditions' of our reactive attitudes. Tognazzini suggests that we should view the facts of moral responsibility as "*at least partly determined by*" our moral practices and reactive attitudes, where the 'partly' leaves room for some kind of external qualifier in the form of propriety conditions. Wallace too presents such a qualified version of the Strawsonian theory:

[27] See (Tognazzini, 2013), cited in (Todd, 2016)

[28] See (Wallace, 1994), cited in (Todd, 2016)

> [T]he facts by reference to which [the question of moral responsibility] is to be decided are specified in terms of our practice of holding people responsible: they are facts about whether it would be appropriate [fair] to adopt towards people the stance of holding them responsible. (Wallace, 1994), cited in (Todd, 2016).

The problem with these various qualifications, however the propriety conditions are cashed out, is that they ultimately undermine the very principle of the Strawsonian reversal – the idea that the explanation of our system of moral responsibility could be "*achieved without postulating a prior and independent realm of moral responsibility facts*" (Wallace, 1994). They amount to what Shoemaker has recently described as a 'shadow scepticism': the idea that the apparently response-dependent properties like 'moral responsibility' (constituted by our reactive attitudes) are in fact shadowing response-independent properties in the world.[29] One way in which this might be cashed out with respect to our moral responsibility (blaming) intuitions is that they are responding to something like 'non-benign norm-violations' – something that, on the shadow-sceptic's account, constitutes a response-independent class of behaviours to which we have simply evolved to be sensitive. As Shoemaker points out though, there isn't a viable response-independent account of either 'non-benign' or 'norm-violating'.

The problem with qualified versions of the Strawsonian reversal, as Todd explains, is similar to the problem of divine command theory.[30] According to divine command theory, something is moral because it is commanded by God, but does that mean that if God had commanded unprovoked murder, then that would be moral? The various qualified versions of the Strawsonian reversal are the equivalent of saying that something is moral because *appropriately* commanded by God, suggesting that God was somehow constrained in terms of what he could have commanded (in a way that seems to imply an exterior realm of moral facts). Saying that moral responsibility is only *partly* determined by reactive attitudes, or that it may be explained in terms of facts about when it would be *appropriate* to hold certain reactive

[29] See (Shoemaker, 2017) for one of the most sophisticated defences of a response-dependent theory of moral responsibility to date, proceeding on the basis of an analogy between moral responsibility and the funny. One way in which I will attempt to modify this account is in avoiding any potentially normative reference to 'properly developed and informed human sensibilities' as the basis of a response-dependent theory. Shoemaker's suggestion is that, subject to differences in 'ecology and social structure' resulting in variable emphasis on different blaming emotions or differences in the expression of such emotions in different cultures, we should be able to reduce concepts like blameworthiness to the fittingness conditions of something like anger. We could call this something like the 'unified response-dependence account': there may not be any unified set of response-independent properties that our properly developed and informed human sensibilities are responding to, but the reactive attitudes have a single set of fittingness conditions based on this form of human sensibility. Shoemaker seems to oscillate between this kind of normative account 'properly developed' human sensibilities and the recognition that "at the end of the day our equally refined sensibilities may just crank out different responses to the same thing." My own account, on the other hand, will try and embrace and make sense of this relativist possibility.

[30] Todd characterises this as an example of a 'Euthyphro-style' relationship in terms of explanatory priority.

attitudes, similarly implies an independent realm of moral responsibility facts and undermines the proposed reversal. The "middle ground", as Todd puts it, "is essentially unstable" (Todd, 2016).

One objection to these 'propriety condition' qualifications to the Strawsonian reversal is that the resulting theory of moral responsibility is not, as Strawson claims, properly incompatible with a libertarian or consequentialist definition of moral responsibility. If the reversal is taken to mean that moral responsibility is determined by facts about when it is *appropriate* to hold certain reactive attitudes, it is be a simple matter to insert the libertarian conditions for moral responsibility (the principle of alternate possibilities) as constituting the 'appropriateness conditions' for particular reactive attitudes, such that the reactive attitudes would simply be 'inappropriate' if the thesis of determinism is true. If, on the other hand, we adopt an uncompromising Strawsonian reversal that makes no reference to propriety conditions, this kind of libertarian approach would be impossible. The libertarian could only succeed if his proposed criteria for moral responsibility (the existence of alternate possibilities) just happened to accurately coincide with people's ordinary responsibility intuitions and reactive attitudes – which, as a matter of empirical fact to be taken up in Chapter III, it would appear that they do not: people are happy to hold agents in determined universes responsible for their actions as long as the causal pathways determining their behaviour operate *through* their deep mental states rather than bypassing them.[31]

All of this raises the question of why theorists find it necessary to introduce these 'propriety condition' qualifiers in the first place. According to Todd, it is in order to avoid the 'absurd conclusion' that if, for example, we were to hold children and the mentally ill morally responsible for their behaviour, then they would in fact *be* morally responsible for it. The first response to this objection is rather simple: to anybody who takes the (unqualified) Strawsonian reversal seriously, the apparently 'absurd' consequence simply does not constitute the kind of 'reductio' argument that proponents of the objection think. Why should it be a problem that, were our moral practices to be different, different things would be right or wrong, or different agents would be considered morally responsible or not? The intuitions that form the basis of the objection come from a perspective well and truly embedded in our own moral practices. The only way that one could reasonably think them to apply them in such a hypothetical scenario, in which the relevant moral practices have been substantially altered, is if one

[31] See, for example, (Knobe & Nichols, 2017; Murray & Nahmias, 2014)

maintained a belief in the existence of moral responsibility facts independent of our moral practices and reactive attitudes. But this, as we know, is a premise that is explicitly rejected by Strawson.[32]

A second response to the objection is to ask whether it is possible, even from within the framework of our own moral practices, to raise doubts about the realist account of moral properties as objective, independent facts. Perhaps the place to start is with the very examples raised in the objection – children and the mentally ill. The realist approach to moral responsibility for agents like these would be to claim that they lack some relevant criterion for morally responsible agency (whether it be the incompatibilist criterion of counterfactual possibility, or indeed the compatibilist criterion of internal sourcehood via deep mental states). Both of these accounts involve representing the facts about moral responsibility as supervening on (objective) properties of the agents in question: in the case of mentally ill agents, either their behaviour is entirely causally determined by their condition such that they are not in fact capable of behaving otherwise than they do, or their behaviour fails to reflect the relevant deep mental states (again, presumably as a result of the condition). These beliefs become more difficult to maintain, however, when we consider a broader range of moral intuitions or reactive attitudes than those blaming attitudes mentioned in the objection.[33]

Take the case of young children: the intuition that the objection is appealing to is that children are not really responsible for their wrong actions because 'they're only children'. Either their behaviour is 'determined' by their basic needs and instincts in a way that rules out moral responsibility, or they simply do not have the self-mastery to express their 'deep' attitudes (if indeed they have them) in action. But what about our *positive* reactive attitudes towards children, where we may attribute positive moral properties to them for which we consider them to be *praiseworthy*? Whilst the situation is admittedly complicated by the fact that we tend to adopt blaming or praising responses towards children regardless of our beliefs concerning their moral responsibility (these might be considered 'contrived' moral responses, for the purposes of moral education) I think we can nonetheless imagine situations in which a child's behaviour might genuinely impress us as evincing praiseworthy deep attitudes. Consider,

[32] Note however that the 'Strawsonian reversal', as I have presented it, is not explicitly defended by Strawson himself, as pointed out in (Todd, 2016, p. 210).

[33] It is a general problem with theories of moral responsibility that responsibility is often reduced to *negative* responsibility (blameworthiness) and reactive attitudes to *negative* attitudes (resentment), such that we do not consult our intuitions on the whole spectrum of reactive attitudes or moral responsibility judgments. A number of authors have noted the 'asymmetry' in our responsibility practices and intuitions, including (Wolf, 1990) and (Nelkin, 2011).

for example, a child's apparently heartfelt attempts to comfort a parent or sibling in distress. We intuitively think that such behaviours show that a child *cares* about the person in question, in a way that elicits genuine moral approval.

Ultimately, what examples like this should show us is that we *don't* think of children as 'incapable' of morally responsible agency. The initial assumption – that children and the mentally ill are not morally responsible for their actions because they lack some property required for morally responsible agency – becomes more difficult to maintain: they seem to be capable of morally responsible agency when it comes to actions of which we approve and yet incapably of it when it comes to actions of which we disapprove. The alternative – that the moral responsibility facts in question are at least partly determined by features of our normative framework – starts to seem a little more plausible. To summarise: the objection is based on the assumption that the fact of, for example, a child's capacity for morally responsible agency is independent of our moral practices because it relies on some objective feature of the agent in question, such that it would be absurd to suggest that a change in our moral practices could change these moral responsibility facts. And yet, a closer inspection of our intuitions surrounding children's moral agency suggests that they might not be quite so independent of our moral practices, sensitive as they seem to be to, amongst other things, our beliefs about what constitutes 'good' behaviour.

To address the very roots of the objection we need to consider the more general form of a persistent and widespread objectivist intuition. Todd's suggestion is that *no* theorist has so far been willing to accept an unqualified version of the reversal.[34] Even theorists prepared to accept that our moral practices play *some* role in defining what it is to be morally responsible also maintain that this relativism cannot be without limits – that there must be some external constraint on what, for any given set of moral practices, can constitute moral responsibility. The intuition itself is worth addressing even if on its own it does not constitute a significant objection to the anti-realist account. It is most plausible that it simply arises because we are unable to fully remove ourselves and our moral intuitions from our own moral context, but there may also be more too it. Consider a more basic kind of moral judgment about what constituted 'good' or 'bad' behaviour. It seems that we are capable of taking a relativist approach to such judgments

[34] Though this would appear to be the position taken by Bennett (1980), to which Strawson was at least generally sympathetic as an interpretation of his own work (Strawson, 1980).

at least to a certain degree.[35] We can imagine that, in different cultures to our own, the rules about what is right or wrong may be slightly different to ours – but our relativism is not absolute. We cannot conceive, I would argue, of a moral practice within which something like unprovoked murder was condoned, and such limitations might naturally lead us to qualifications of the anti-realist position like the various propriety conditions mentioned above.

I would argue, however, that these limitations on the conceivable range of moral possibility are best understood not as evidence of set of response-independent moral facts, but as a recognition of a kind of evolutionary or developmental constraint operating on moral practices generally. It may simply strike us as impossible or implausible that such a moral practice – where, for example, unprovoked murder is condoned – could ever develop or sustain itself. When we consider moral practices in a functional context, considering their role in maintaining community bonds and promoting prosocial behaviour, it must strike us that certain practices simply never could exist, or would at least be extremely unlikely to exist. Were we to regularly hold individuals responsible for behaviours (reproaching them, punishing them) that, as a matter of fact, our interventions had no prospect of ameliorating, then we might imagine a sense in which the practice in question is *maladaptive*. Do such intuitions require us to assume the existence of underlying, response-independent properties that our reactive attitudes and the corresponding moral concepts are tracking? According to Shoemaker this form of evolutionary shadow scepticism is unlikely to reveal any properly response-independent properties:

> The anger that we have in response to certain actions and attitudes is a function of what we care about, which is indeed a function of our specific biological and cultural history, our anthropological nature. The shadow skeptic's claim, though, seems to be that we evolved to care about norm violators and so evolved to get angry with them because doing so promoted our ancestors' reproductive success. That is to say, there was a (response-independent) property—nonbenign norm violator—that some agents responded to with anger, where doing so somehow increased their fitness, so that we, their descendants, inherited that response mechanism. But this would be quite incredible, if true. Rather, what seems more in line with the science on this score is that our early ancestors responded with anger to some people because doing so made their targets behave in certain ways and made the angry people themselves behave in certain ways (so as to avoid the anger of others or provide some prudential reputational benefit), and such behavior promoted their reproductive success, so was inherited by us. On this latter explanation, there is no tracking of response-independent properties being done at all. Indeed, bolstering this explanation is the fact that it is entirely unclear just how the (response-independent) fact of something's being a nonbenign norm violation per se could make anger at the violator reproductively advantageous. (Shoemaker, 2017, references omitted)

[35] This has been empirically confirmed by Sarkissian and colleagues (Sarkissian, Park, Tien, Wright, & Knobe, 2014) contrary to the general assumption that people's moral judgments are always subject to an objectivist intuition.

Essentially, the idea that some moral practices might be more 'adaptive' than others does not commit us to the idea that the reactive attitudes that constitute those moral practices are tracking response-independent properties. What is adaptive will inevitably be a (response-dependent) feature of various contingent facts about all of the agents involved.

What does all of this mean for deep selves? What would an account of the deep self concept look like on a proper (unqualified) formulation of the Strawsonian reversal? We need to begin by considering the role that the deep self might play in the kind of qualified Strawsonian reversal considered above, in which what constitutes morally responsible agency is determined by the propriety conditions of our reactive attitudes. Within Strawson's own account the reactive attitudes are understood as responses to perceived 'quality of will' in other agents, such that the kind of explanation required to mitigate our reactive attitudes is one that convinces us that the agent towards whom those attitudes are directed does not in fact have the relevant quality of will. A simple way of transforming this into response-independent theory of moral responsibility is to say that a moral responsibility judgment is correct just when the object of that judgment in fact possesses the relevant quality of will. As an example, resentment would be appropriately directed towards an agent whose behaviour manifests *ill will* towards me and so the judgment that she is blameworthy for that behaviour would be correct. What this 'propriety conditions' account relies on is there being an independent fact of the matter about an agent's quality of will – on 'quality of will' being a response-independent property. This is essentially what realists about the deep self seek to provide: a response-independent property capable of filling in this picture and providing the propriety conditions that *justify* our reactive attitudes by reference to independent moral responsibility facts.

The alternative, anti-realist approach – based on an unqualified formulation of the Strawsonian reversal – is exactly what I will seek to develop. It does not deny that reactive attitudes are responses to perceived quality of will, but it denies that there can be anything more than *perceptions* of qualities of will. There can be no fact of the matter as to whether somebody *really* bears us good or ill will – no independent facts of the matter for the kind of question that our deep self attributions seek to answer. This is not of course to deny that there are regularities in different people's behaviour – reliable behavioural dispositions on the basis of which we regularly make successful predictions. But these regularities are fundamentally different to the subject of our deep self intuitions: our deep self attributions towards a particular agent are not concerned merely with predicting her future behaviour. The willing and unwilling addicts – two classical examples for deep self theorists – both manifest the same behavioural dispositions

despite the fact that we explain the behaviour of one as expressive of her deep self and the behaviour of the other as contrary to it.

The idea that deep attributions are not tracking real, response-independent properties in the agents towards whom they are directed – the idea, indeed, that there is no objective reality to what we might describe as 'quality of will' – is likely to be controversial. One way of conceptualising the deep self that we are so used to referring to as a real property of agents might by analogy to something like 'secondary qualities' described by Locke and later by Hume. Whereas primary qualities like size and shape play deep explanatory roles in the behaviour of objects independently of our 'sensory apparatus', secondary qualities like colour have explanatory roles that only feature in their interactions with us. Colour is perhaps a particularly apt analogy: whilst there is certainly a fact of the matter about the different wavelengths of light that a surface is disposed to reflect, a claim that an object is 'red' (i.e. produces particular sensations) can only ever be true in the context of a particular sensory apparatus and a particular constitution of ambient light. An agent with a different sensory apparatus would disagree, as would an agent viewing the object under different lighting conditions, and there would be no way of objectively resolving the dispute. On the account of the deep self that I am proposing, there is certainly a fact of the matter about how a given agent behaves, but this alone is incapable of verifying or falsifying any particular claim about that agent's deep self. It is only in the context of a particular sensory apparatus (a particular folk-psychological cognitive mechanism) and particular background conditions (models of human behaviour informed by moral norms and salient social narratives) that a claim about an agent's deep self can be true or false.

Chapter II: The Realist Assumption in Philosophical Theory

§1) Introduction

In the previous chapter I outlined the concept of the 'deep self' that is at work in our folk psychological and moral practices. I set out the scope of the question concerning its nature and its place in our attributions of moral responsibility, highlighting the central role that it seems to play in such judgments. In confining the scope of my enquiry to compatibilist approaches to moral responsibility, I set out two possible alternative conceptions of the deep self. On the more traditional, 'realist' approach, the deep self is an objective (response-independent) psychological property of agents, which explains and justifies our moral responsibility attributions towards them. On the alternative approach, which I have described as involving a 'Strawsonian reversal', the deep self, or rather 'deep self properties', are instead response-dependent. This would mean that there is no objective property – no 'fact of the matter' of moral responsibility – that all 'correct' moral responsibility judgments are accurately tracking. On this anti-realist approach, the deep self concept itself plays a pivotal role in the construction of fundamentally response-dependent moral responsibility 'facts'.

My aim in this chapter is to provide an overview of a range of compatibilist philosophical approaches to the deep self and its role in our moral responsibility judgments. The survey is by no means comprehensive, but an analysis of some of the major accounts reveals a general point about the realist assumption underlying a much broader range of deep self theories. Though the theories I will consider do not necessarily reliably use the terminology of 'deep selves', what they all have in common is an attempt to identify a property – a 'compatibilist criterion of internal sourcehood' – capable of explaining and justifying our patterns of moral responsibility attribution. They are all also at least implicitly committed to the realist assumption set out above: that there are response-independent facts (a unified set of psychological properties, capable of objective description) constitutive of the deep self that our (correct) moral responsibility attributions are responding to. A consequence of this assumption is that these approaches also all focus entirely on 'agent-centred' factors – the mental states, dispositions, or capacities of candidates for responsibility attribution – whereas my analysis will aim to show why a wider lens is needed in order to fully explain our patterns of attribution.

Within these compatibilist, realist approaches to the deep self there is significant disagreement as to how exactly the relevant set of psychological properties are to be

characterized. In addition to the distinction outlined in Chapter I (between those accounts that take the relevant mental states to be a subset of desires and those that take them to be a subset of beliefs/evaluative judgments) there are a range of claims about the importance of conscious, reflective endorsement, causal history and continuity, and various other factors potentially singling out the relevant 'deep' psychological elements. As I argue for the inability of these factors to sustain the realist assumption, I will also come to question the principal distinction between characterisations of the deep self in terms of desires or evaluative judgments.

Whilst various theorists may have come across factors that seem to make a deep self 'story' more plausible, I will argue that none is yet immune to clear counterexamples, and that ultimately none puts forward a robustly response-independent criterion. In order to show this, I will begin by looking at the Frankfurtian (desires-based) tradition in depth, including some relatively recent contributions [§2] before considering the Watsonian (evaluative judgment-based) tradition to argue that it suffers from extremely similar flaws [§3]. What emerges from an analysis of these approaches is not only their inability to support the realist assumption with any kind of response-independent elements constituting the deep self, but also the emergence of some interesting, response-*dependent* factors that seem to be at work in our moral responsibility judgments (to be examined in more depth in Chapter III). Ultimately I will argue that the deep self is not so much a psychological reality, capable of explaining and justifying our attributability intuitions, but a flexible placeholder in our moral explanations – a *folk*-psychological construct – serving to rationalize them. Let us begin, however, at the beginning.

§2) The Frankfurtian Approach

Harry Frankfurt's approach to the deep self, beginning with his defence of a compatibilist concept of moral responsibility, is an important foundation for all subsequent contemporary approaches to the deep self. It begins with a rejection of the (incompatibilist) 'Principle of Alternate Possibilities' – the idea that an agent is only morally responsible for her actions in situations in which she *could* have acted otherwise than she in fact did (Frankfurt, 1969). The argument proceeds from a number of 'Frankfurt-style cases': cases where an agent fully identifies with one course of action[36] ('of her own free will') whilst, unbeknownst to her, some

[36] What is meant by this 'identification' will be the subject of significant debate later on, but for now it is sufficient if we understand it according to its ordinary (vague) meaning, as suggesting that the agent consciously endorses, embraces etc. the relevant course of action, whatever we understand such terms to mean within our folk psychology.

external influence ensures that, whatever her will, she will in fact end up taking that course of action. Her decision in such cases is *overdetermined*: it is determined both internally, by the exercise of her free will, and by some external influence. What we find in such cases is a strong intuition that the agent in question *is* responsible for her actions, even though she is not, in fact, capable of doing otherwise.

The next step, having rejected the incompatibilist criterion for morally responsible agency, is to come up with a compatibilist alternative. The resulting 'internal sourcehood' criterion for moral responsibility (Frankfurt, 1971) is, in its general form, the foundation of all subsequent (compatibilist) accounts of the deep self and morally responsible agency. Whilst Frankfurt's account is cashed out, at least initially, in terms of a concept of 'personhood', this should not hide the fact that he is, from the beginning, dealing with the same concept described by Strawson as 'freedom'.[37] What both authors are trying to capture are "those attributes which are the subject of our most humane concern with ourselves and the source of what we regard as most important and most problematical in our lives" (Frankfurt, 1971) – in other words, the source of morally responsible agency. For Frankfurt, this is to be found in what he describes as 'the structure of a person's will' – or, in more everyday terms, the particular way in which we *want* things.

Whilst a whole range of beings are capable of *wanting* things (in the way, for example, that we might describe a dog as 'wanting' a bone) we would not describe all such beings – or all such instances of 'wanting' – as manifesting morally responsible agency. Clearly the concept of 'wanting' alone is problematically vague, and incapable of telling us anything important about moral responsibility attribution. As Frankfurt points out, a statement of the form '*A* wants to *X*' could be consistent with any number of different psychological states:

> Such a statement may be consistent, for example, with each of the following statements: (a) the prospect of doing X elicits no sensation or introspectible emotional response in A; (b) A is unaware that he wants to X; (c) A believes that he does not want to X; (d) A wants to refrain from X-ing; (e) A wants to Y, and believes that it is impossible for him both to Y and to X; (f) A does not "really" want to X; (g) A would rather die than X; and so on. (Frankfurt, 1971)

[37] The move away from the terminology of 'freedom' in general is not insignificant. Many compatibilists would argue that moral responsibility is compatible with the truth of determinism *in the absence* of free will, in the most demanding sense of the word implying counterfactual possibility. Alternative compatibilist accounts may claim that not only moral responsibility but also this more demanding kind of free will are both compatible with determinism, but it is not with claims of this latter kind that this thesis is concerned. For our purposes it will be important to bear in mind the special sense in which Strawson and others after him use the word 'freedom': not to refer to a traditional concept of freedom, but rather to describe 'being constrained *in the right way*' – in the way necessary for morally responsible agency.

In order to improve upon the rather unhelpful concept of 'wanting', to discover what it is about *some* of our desires that is relevant to morally responsible agency, Frankfurt introduces some new terminology. **First-order desires**, within Frankfurt's framework, are desires that have as their object a particular course of action – desires *to X* – without implying that such desires will necessarily manifest themselves in any observable degree in an agent's behaviour. First-order desires are elements of an agent's psychology that "*merely [incline] an agent in some degree to act in a certain way*" _. One may be unaware of certain of one's first order desires, or have conflicting or simply incompatible first-order desires. Amongst these first-order desires, the one on which an agent ends up acting – the desire that is 'effective in action' – is deemed to be the agent's **will**, where this is distinct from an agent's plans or settled intentions (on which an agent may fail to follow through if he turns out to have a stronger, conflicting desire). **Second-order desires** are desires that have as their object desires of the first order (a 'higher order' desire is one that has as its object a desire of a lower order), meaning that they are desires to have certain desires. Where these second-order desires are for a particular first-order desire to be effective in action (instead of, for instance, desires simply to experience a particular desire – consider Ulysses' desire to feel the pull of the siren song without wanting to actually be moved to throw himself onto the rocks) then they are referred to as **second-order volitions**. The difference between morally responsible agency and other forms of action to which we do not attach moral responsibility is, according to Frankfurt, the presence of second-order volitions. Agents are morally responsible for just those actions that reflect their higher order volitions.

In the context of Frankfurt's compatibilist project (of explaining how we may attribute moral responsibility to an agent who is not capable of doing otherwise than she in fact does) this relationship between actions and second-order volitions constitutes the 'internal sourcehood' criterion. We correctly attribute moral responsibility to an agent for a given course of action when her will is structured towards it in this particular way, so that her action *expresses* her higher order volitions. 'Freedom of the will' in this compatibilist framework (as opposed to simple freedom of action) is a freedom that only morally responsible agents may possess – indeed that is constitutive of morally responsible agency. Whilst freedom of action, understood as the freedom to translate first order desires into actions, may be restricted by any number of circumstances external to an agent, what is much more relevant with respect to moral

responsibility is the range of ways in which an agent's freedom of will may be impeded.[38] This would typically involve *internal* (psychological) impediments to the translation into action of particular desires in accordance with an agent's second order volitions.[39] Frankfurt illustrates this with the example of the 'unwilling addict', whose (first order) desire for a drug is so strong that it inevitably ends up moving him to action (taking the drug) even though his higher order volitions speak against it. This is an example of a desire from which an agent is *alienated* (as opposed to ones with which she *identifies*, through the formation of higher-order volitions). According to Frankfurt, it explains why *some* desires are attributable to agents whilst others are not: desires with which one *identifies* are understood to be internal to an agent's deep self and so properly attributable to that agent for the purposes of our moral responsibility judgments, whereas it would be a mistake to attribute moral responsibility to an agent for a desire from which she is alienated.

From this starting point, a number of significant objections to the Frankfurtian account have been raised, in response to which it undergoes significant developments. These developments are the subject of the following sections, where I will argue that the account of the deep self that we are left with as a result of them is entirely incompatible with the realist assumption that the deep self tradition takes as its starting point.

(i) The Objection from Manipulation Cases

One early objection to Frankfurt's account of the deep self (which in fact applies more generally to any account of moral responsibility that identifies an agent's essential self for the purposes of attribution with a particular set of mental states) is the manipulation objection.[40] The objection comes in a range of different forms, from interventions by evil scientists or post-hypnotic suggestion to 'Brave New World'-style brainwashing or even simply unfortunate

[38] There are of course many 'excusing conditions' relevant to our moral responsibility judgments, including non-culpable ignorance, accidents, etc. What is different about the kind of excusing conditions we are interested in here is that they serve to excuse agents for *intentional* actions. Whereas an agent who accidentally steps on my foot was never aware that she was acting under that description ('stepping on my foot' as opposed to merely 'walking') an agent who acts, for example, out of compulsion *is* aware of the (morally significant) description under which she is acting. The way in which she is excused is therefore fundamentally different.

[39] 'Internal', in the sense used here, simply means part of an agent's psychology, so as to distinguish 'psychological' impediments to acting on a given desire imposed by the objective reality of the outside world. A desire for *x* may never move me to action because of a stronger and incompatible desire for *y*, but this is very different from a desire for *x* never moving me to action because the necessary conditions for *x* never obtain in the world. Later the internal/external distinction will be used by deep self theorists to draw a different distinction: between those elements of an agent's psychology located *within* the deep self and those *outside* it.

[40] See, for an in-depth formulation of the objection, (Fischer & Ravizza, 1998, pp. 194–206)

formative circumstances. What they have in common is the following: some mechanism (imagine any of those listed above) is used to manipulate a certain element of an agent's psychology (in the Frankfurtian case a higher-order volition) concerning some particular course of action. According to Frankfurt's original theory, what is relevant for the purposes of moral responsibility is the relationship of one's higher-order volitions to one's will, and so it shouldn't matter how an agent came to have a particular higher-order volition. And yet, when faced with cases of, for example, post-hypnotic suggestion, we intuitively think that the agents in question are *not* responsible for their actions, even though they do have the relevant higher-order desire. History somehow seems to matter.

One potential resolution to this problem might be found in Watson's later distinction between responsibility-as-attributability and responsibility-as-accountability, as outlined earlier in Chapter I (Watson, 1996). As Watson observes, our responses to manipulation cases are often not simply the withdrawal of our moral judgments but rather a kind of deep ambivalence. This he explains in terms of on the one hand maintaining our attributability judgments (that the actions are properly attributable to the agent's deep self: she is, at her core, a bad person, in the sense of identifying with or endorsing bad actions) whilst on the other hand withholding accountability judgments, either because the agent appears so deformed as to not be an appropriate subject of our moral demands or (as Watson earlier suggests) because we do not feel as though we have the *standing* to criticise someone in her situation, not knowing how we might have turned out in the same circumstances. According to Watson, it is the conflicting demands of these two intuitions (to find responsible but not to hold responsible) that account for our ambivalence when faced with such cases.

The ambivalence in question however, whilst strong with respect to more ordinary, 'causal history' examples used by Watson,[41] is by no means uniform across the full range of manipulation cases. Our intuitions with respect to the more far-fetched 'sudden intervention' manipulations are noticeably less ambivalent, as we tend to see an agent acting, for example, under the influence of some sinister psychological manipulation, as more completely exculpated by those facts. To a certain extent this may be due to the implausibility of the

[41] An example of 'manipulation by ordinary causal history' is the case of Robert Alton Harris, a remorseless serial killer and the product of an extremely abusive childhood (Watson, 1987b). On the one hand his complete lack of remorse and apparent identification with/endorsement of his crimes leads us to attribute moral responsibility for them to him, and yet when we hear the detail of his unfortunate childhood we are left with an incoherent set of responses, torn between holding him responsible for who he is and exculpating him on the basis of how he came to be like that (causal history).

examples in question, as we are more likely to understand the changes in question as 'superficial': an agent may be moved to act in a certain way by a particular, manipulated, higher-order volition, but there remains an intuitive sense in which she still does not *really* want to do it. Arguments to this effect however (that the higher-order volition is not really hers; that it is not properly 'integrated' into her psychology; or other similar explanations) fail to engage however with what is stipulated by the thought experiment: that the higher-order volition in question *is* hers.

Manipulation cases can be constructed around any particular concept of the deep self such that, by stipulation, the deep self of the manipulated agent is not bypassed – for example, the manipulation involves creating within the agent a particular mental state that appears to satisfy the relevant description of a deep mental state – and yet people typically judge such agents not to be morally responsible for their actions. Anyone who endorses Frankfurt-style conditions on moral responsibility ought perhaps to reject such a theory when faced with manipulation cases, but what appears to be a more common folk-psychological response is to attempt to reconcile the deep self theory with the particular responsibility intuition in a manipulation case – finding ways in which, despite the stipulations of the manipulation case, the deep self *is* ultimately bypassed – revealing a degree of flexibility in the underlying folk theory.

What is interesting about these resistant intuitions is that they reveal to us the influence of a phenomenon known as 'psychological essentialism'. Though the phenomenon will be explored in more detail in Chapter III, the general idea is that humans have a tendency to view certain kinds of entities as having 'essences' – underlying natures that are causally responsible for their apparent features – and to think of these essences as persisting across 'superficial' changes to the entity in question. In this case, we seem to believe that the *essence* of the agent in question – her deep self, perhaps – remains unchanged across the manipulations described in the thought experiment, leading to the judgment that her deep self has been 'bypassed' by the causal pathways that lead to the action in question.[42] Even returning to the 'ordinary causal history' cases, it is not impossible that the essentialist intuition persists. It can be seen in

[42] Much interesting empirical work has been done on the topic of folk intuitions concerning compatibilism, with an important result emerging from the research being that people typically (mis)understand the thesis of determinism to involve the *bypassing* of an agent's relevant psychological states (deep self) by a causal chain leading to her action. When the thesis of determinism is explained so as to rule out the possibility of bypassing – making it clear that the causal chains that determine an agent's action operate *through* that agent's psychological states – incompatibilist intuitions more or less disappear. For some of the initial empirical results, see (Nichols & Knobe, 2007) & (Phillips & Knobe, 2009); for the explanation in terms of bypassing, see (Murray & Nahmias, 2014) & (Knobe & Nichols, 2017). See also Chapter III.

statements like that of Watson, describing our intuitions towards Robert Alton Harris as involving "sympathy towards the boy he was […] at odds with outrage toward the man he is" (Watson, 1987b). Our moral responsibility attributions reveal a belief that the boy he was somehow persists underneath the murderer he has become. It seems that plausible deep self story must be grounded in some kind of continuity: in order to attribute responsibility to an agent in the present for behaviours in the past there needs to be continuity between the agent's present and past self – a kind of continuity that the deep self concept appears to provide.

What this means for Frankfurt's account of morally responsible agency is that, whilst higher-order volitions may still appear relevant, they are not the whole story. History seems to matter too, in the sense that attributing some attitude or moral quality to an agent's deep seems to carry with it the implicit judgment that the agent has always (deep down) been that way. The relevance of continuity to any plausible deep self story is thus at odds with the Frankfurtian account of the 'formation' of higher order desires (through a process of reflective self-evaluation). Frankfurt's account would allow an agent's deep self to change from one moment by the next, reconstituted by a mental act of 'identification', and so needs to be modified in order to capture the judgment of continuity involved in deep self moral responsibility attributions. But does our essentialist folk-psychology and the continuity it imposes rule out the possibility of responsibility attributions following 'Road to Damascus' conversions? I would argue that it doesn't, but only because we are able to interpret such changes not as fundamental alterations to the content of the deep self but rather as *revelations* (to the agent and potentially to others as well) of what an agent's deep self contained all along.[43] In this sense, part of the role of the deep self concept in responsibility attributions is to impose the necessary continuity on the agents in question. Moreover, it is this sense of revelation of an underlying truth – of continuity – that is distinctly missing from the manipulation cases, where we are explicitly presented with an account of manipulation (from without) *as opposed to* expression (from within). There are exceptions to the rule that our moral responsibility attributions always require – or impose – a continuous deep self. Traumatic head injuries are such an example: we may attribute responsibility to an agent for who she is after such an injury,[44] but the 'manipulation' involved in the head injury is such that it severs the continuity between the person she was and

[43] Perhaps the most appropriate analogy is with the declaratory theory of the common law, according to which judges never actually 'change' the law but rather reveal or declare what it has always been. See e.g. (Beever, 2013). This also seems to be the case even in the original 'Road to Damascus' conversion, which is most often interpreted as involving Paul 'seeing the light' as opposed to merely 'changing his mind'.

[44] Consider, for example, the case of Phineas Gage, as described in (Tobia, 2015)

the person she is, such that we can both regret the loss of the former and make responsibility attributions towards the latter. The idea, however, of the deep self being suddenly altered, or even constituted, by a mere mental act of identification conflicts with our intuitions about moral responsibility attribution, which seem to require a degree of continuity.[45] Whatever story ends up being told about the structure of the will or higher order desires will need to be modified to reflect this.

(ii) The Hierarchy Objection

A further objection to Frankfurt's account of the deep self is formulated in Watson's 'Free Agency' (Watson, 1975) and comes in two parts. First, if what Frankfurt's account presents is a 'hierarchy' of desires, where does the progression to higher order desires end? Second, what is it about a particular kind of desire – i.e. one that has another desire as its object – that gives it priority over any other desires for the purposes of attributability? If a higher-order desire may always be formed, what is it about any *particular* desire that gives it priority over the others? "Since second-order volitions are themselves simply desires, to add them to the context of conflict is just to increase the number of contenders; it is not to give a special place of any sort to those in contention" (Watson, 1975, pp. 217–218).

Frankfurt's pre-emptive response to this objection is that an act of *decisive identification* (the outcome of the process of self-reflective evaluation referred to above) has the effect of removing the possibility of conflict at higher orders and giving the relevant priority to the higher-order volition in question. But Watson insists that the appeal to a notion of decisive identification *is* arbitrary, and cannot explain why higher-order volitions have the special property of constituting one's deep self for the purposes of moral attributability, prompting Frankfurt to provide a more detailed account of the notion of 'identification' (Frankfurt, 1987).

The analogy that Frankfurt presents to explain 'decisive identification' is between deciding which desires one wants to be one's will and attempting to solve a problem in arithmetic. Framing Watson's objection in terms of the latter case, he asks how it is that the mathematician can be satisfied with a given answer: he may repeat the calculation any number of times, but it is also possible that the same mistake be repeated any number of times, so how

[45] This also seems clear when we reflect on the nature of moral responsibility judgments: they only make sense if there is a degree of continuity in an agent's life such that there is a link between a past act and the present person to whom we attribute responsibility for it. Concepts like the deep self, or 'character', etc., are employed to make that link. See e.g. (George Sher, 2005, p. 18).

can the repetition of calculations be stopped at a point that is non-arbitrary, meaning that the mathematician identifies himself with the result? Whilst it is possible, Frankfurt suggests, that such a sequence be terminated because of mere laziness, or because the mathematician is otherwise distracted (in which case he fails to identify himself with, or endorse, the answer at all), it is also possible for the mathematician to decide for a reason to end the sequence – specifically, for the reason that he is confident that his answer is correct (or at least confident enough that he would consider further calculations to be a waste of time). In terms of desires:

> [A] person may be led to reflect on his own desires either because they conflict with each other, or because a more general lack of confidence moves him to consider whether to be satisfied with his motives as they are. [An agent] can without arbitrariness terminate a potentially endless sequence of evaluations when he finds that there is no disturbing conflict, either between results already obtained or between a result already obtained and one he might reasonably expect to obtain if the sequence were to continue. (Frankfurt, 1987)

The calculation involved in reflective self-evaluation involves an analysis of one's own desires with the goal of establishing whether they are in fact desires by which one wishes to be moved. The calculation at the level of first-order desires would simply involve a comparison of the strengths of various competing first-order desires, but a person unsatisfied with the result may engage in a second calculation (the formation of second-order desires) to determine if what she most wants, as determined by the relative strength of her first-order desires, is *really* what she most wants. Identification with a desire is decisive when one's conclusion at one given level (e.g. at the level of second-order desires, where the conclusion consists in the formation of second-order desire) leaves no room for further doubt or conflict. But ultimately the analogy with a mathematical calculation proves too fitting. In arithmetic, there simply *is* a correct answer. To stop after any particular number of calculations may not be entirely arbitrary – in the sense that the agent may have attained a reasonable degree of confidence in the conclusion – but it does not *make* the answer arrived at the correct one. The correct answer is simply a matter of fact, over which the agent's (conscious) identification with or endorsement of any particular answer can have no influence.

Ultimately this is the conclusion that Frankfurt comes to with respect to the deep self as well. The concept of an act of identification is stripped of any relevance it might have had to the question of moral responsibility attribution. What matters instead is whether or not the act of identification is *felicitous*, in the sense that the desires with which an agent consciously identifies herself are in fact those desires that, independently of that conscious identification, happen to constitute her deep self. As Frankfurt explains:

> A person may fail to integrate himself when he makes up his mind, of course, since the conflict or hesitancy with which he is contending may continue despite his decision. All a decision does is to create an intention; it does not guarantee that the intention will be carried out. This is not simply because a person may always change his mind. Apart from inconstancy of that sort, it may be that energies tending toward action inconsistent with the intention remain untamed and undispersed, however decisively the person believes his mind has been made up. The conflict the decision was supposed to supersede may continue despite the person's conviction that she has resolved it. In that case the decision, no matter how apparently conscientious and sincere, is not wholehearted: Whether the person is aware of it or not, he has other intentions, intentions incompatible with the one the decision established and to which he is also committed. This may become evident when the chips are down and the person acts in a way ostensibly precluded by the intention on which he thought he had settled. (Frankfurt, 1987)

The point Frankfurt makes is that whilst an agent can change his mind, he cannot change his deep self. Whether or not a change of mind results in 'wholehearted' identification depends on whether or not the state arrived at coincides with the independent reality of the deep self – independent, that is, of the agent's will or conscious identification. The newly introduced concept of 'wholeheartedness' thus has little if anything to do with the act of identification itself, but is merely a way of designating those higher order desires that we would attribute to an agent's deep self. The idea of a mental act of self-determination is thus abandoned altogether. As put by Frankfurt in his 1992 Presidential Address:

> A person cannot make himself volitionally determinate, and thereby create a truth where there was none before, merely by an "act of will". In other words, he cannot make himself wholehearted just by a psychic movement that is fully under his immediate voluntary control. The concept of reality is fundamentally the concept of something which is independent of our wishes and by which we are therefore constrained. Thus, reality cannot be under our absolute and unmediated volitional control. The existence and the character of what is real are necessarily indifferent to mere acts of our will. Now this must hold as well for the reality of the will itself. A person's will is real only if its character is not absolutely up to him. It must be unresponsive to his sheer fiat. It cannot be unconditionally within his power to determine what his will is to be, as it is within the unconstrained power of an author of fiction to render determinate – in whatever way he likes – the volitional characteristics of the people in his stories. (Frankfurt, 1992)

Though Frankfurt does not explicitly use the language of 'deep selves', it is clear that he is committed to there being a fact of the matter about what an agent's will really is – a deeper reality with which an agent's conscious assessments or identifications may or may not coincide. Whilst this deep self concept necessarily plays a role in Frankfurt's theory of wholehearted identification, it no longer has the same specific content in terms of 'higher-order desires' as Frankfurt's initial concept. The idea of intentional 'formation' of higher-order desires, or of decisive (intentional) acts of identification has essentially been abandoned, though the underlying concept of a deep self has been retained without any specific form or content.

The move away from self-determination is ultimately inevitable for a sourcehood-compatibilist theory of moral responsibility, though it has taken some time for the full implications of the determinist thesis to be acknowledged. Compatibilism necessarily involves holding agents morally responsible for what they do and who they are even when those things are entirely determined in a way that is strictly speaking beyond their control. Though the sourcehood compatibilist can readily accept that an agent has guidance control over some of her actions, in the sense that those actions flow in the right way from her deep self, it is difficult to imagine a sense in which an agent could possible have control over the *content* of her deep self as the concept of self-determination originally presented by Frankfurt would require. As such, a compatibilist theory of moral responsibility centred around the deep self must come to terms with holding agents responsible for who they are (their deep selves) when this is in no way under their volitional control.[46] The concept of self-determination – that is, determining the content of one's deep self, as opposed to merely exercising guidance control over one's actions – makes little sense in a sourcehood-compatibilist context. Having removed the concept of decisive identification as self-determination from the equation, the difficult question that still faces Frankfurt and others is how to differentiate those desires that constitute an agent's deep self from those more superficial desires that are not attributable to her. Without the possibility of recourse to a decisive act of *self*-determination to confer that special status on certain of an agent's desires, there must instead exist an independently and antecedently privileged subset of mental states if the realist approach to the deep self and moral responsibility is to remain viable. It is with claims of this nature that the next sub-section is concerned.

(iii) Dissecting the 'Cares-Based' Approach

The objections canvassed above force the Frankfurtian theorist to accept two things about the deep self: (i) that the content of the deep self is not necessarily reflected in our conscious endorsement of particular desires; (ii) that the deep self is an independent, external reality – it may be revealed to us in particular moments or by acts of reflective self-evaluation, but it is not constituted in such moments or susceptible to change by such evaluation. For those who remain committed to the idea of a particular class of 'deep' desires relevant to morally responsible

[46] The absence of control in question goes rather deep: it is not only the thesis of determinism that precludes an agent's 'doing otherwise' – in the sense of choosing to be something other than what she is – but a more basic logical problem: if we accept that there can be some kind of control over an agent's deep self, what might plausibly exercise such control without involving a regress into deeper and deeper selves?

agency – what Frankfurt labels 'cares' (Frankfurt, 1982) – the question then becomes how to identify them.

According to Frankfurt, the existence of 'cares', as elements of the relevant subset of desires, is something we can *see*, manifested in certain kinds of behaviour. To show this, he presents cases of 'volitional necessity' – cases where an agent finds herself forced, by some *internal* compulsion, to take a particular course of action, even one she had consciously decided against. The paradigm example is Martin Luther's 'Here I stand; I can do no other' (Frankfurt, 1982) but other cases include, for example, the mother who intends to abandon the child she knows she cannot afford to keep but finds she cannot do it, or the man driven to extremes of desperation who still finds himself unable to consume human flesh even despite his resolve that he must do so in order to survive (Frankfurt, 1988). It is suggested that these cases reveal something fundamental about an individual's nature – something central to her identity as a morally responsible agent:

> To the extent that a person is constrained by volitional necessities, there are certain things that he can't help willing or that he cannot bring himself to do […] The essential nature of a person consists in what he **must** will. The boundaries of his will define his shape as a person. (Frankfurt, 1998, emphasis mine)

The difficulty, however, is in distinguishing what Frankfurt describes as 'volitional necessity' from other ways in which an agent's will may be constrained – described by Watson as the various 'volitional **necessities**' (Watson, 2004, emphasis mine). As Watson points out, one may just as well find oneself forced to (or unable to) take a given course of action as a result of "*sources of motivation whose force is independent of one's endorsement or of what one cares about*" (Watson, 2004b, p. 122) – referring to what we might otherwise call 'compulsion'.[47] 'Cares-based' volitional necessity and simple cases of compulsion share essentially the same structure. Consider relatively straightforward cases of depression or various phobias: these can involve agents intending to take certain courses of action (to get out of bed, to meet new people, etc.) consciously endorsing those courses of action and identifying with them (in whichever terms this is cashed out) and finding themselves ultimately completely

[47] Ultimately Watson's argument focuses on the divergence of caring from endorsement, arguing that endorsement (in the sense of conscious evaluative judgment) is more central to agency than cares, whose apparent 'volitional necessity' can be overcome by evaluative judgment. Whilst claims of this nature are interesting in terms of the conflict between Frankfurtian and Watsonian traditions (outlined in Chapter I, and to be taken up again in the following section), what is far more interesting is Watson's observation that there are alternative potential sources of 'necessity' – including many we would consider 'external' to the self – forcing us to wonder how exactly the distinction between 'deep', internal necessity and superficial, external compulsion is being made.

incapable of following through. And yet we certainly would not say of a depressed person that her inability to get out of bed is attributable to her 'deep self' or her fundamental cares.

The realist's approach when faced with these two different kinds of volitional 'incapacity' – one of which appears to be reveal something about or be expressive of the agent's deep self and the other of which does not – is to insist that there must be something different between the two scenarios, and to try and find some (response-independent) factor capable of differentiating 'cares-based' volitional incapacity from compulsion. The idea of a special kind of internality, with 'cares' as the set of motivational factors located deep *within* an agent, appears to provide such a difference-maker. 'Cares' would contrast with those factors imposed on an agent from without, for example as a result of motivational 'pathologies' like depression, compulsions, phobias, addiction, etc., and, if they actually exist as response-independent properties of individual agents (if the dividing line between 'internal' and 'external' within an agent's psychology represents some kind of real psychological distinction), justify the realist assumption. The Strawsonian (response-dependent) theorist's approach, it is worth remembering, would be to start with the observation that *we*, as observers, attribute one set of actions to an agent's deep self but not another, and to ask what might be different about the two different scenarios *taken in their entirety* (that is, including features of the observer – a view that will be developed in more depth in Chapter III). For the moment however, as we are engaged in evaluating the realists' approach to the problem, we must zoom in on the agent's psychology, to see whether the line drawn to separate 'internality' from 'externality' is anything more than a line separating those desires we take to be attributable from the rest; whether the distinction between compulsion and 'cares-based' volitional necessity exists independently of our lens as observers, embedded in a framework of normative expectations.

On this front, the realist accounts are elusively vague. Frankfurt begins by describing a sense in which cares play a 'guiding role' in an agent's life. Unlike mere desires or beliefs, the notion of caring is supposed to "*[imply] a certain consistency or steadiness of behaviour*" – but this comes with some immediate qualifications:

> It is not to be presumed, of course, that whenever a person's life displays over a period of time some more or less stable attitudinal or behavioural disposition, this reflects what the person cares about during this time. After all, patterns of interest or of response may be manifestations only of habits or of involuntary regularities of some other kind; and it is also possible for them to develop merely by chance. (Frankfurt, 1982, pp. 82–83)

Certainly the behaviour of the unwilling addict or the depressive agent exhibits a significant degree of consistency, so something more must be meant by the 'guiding role' that cares play in a life. This is what Frankfurt appears to be getting at when he claims that

> The moments in the life of a person who cares about something […] are not merely linked inherently by the formal relations of sequentiality. The person necessarily binds them together, and in the nature of the case also construes them as being bound together, in **richer ways**. (Frankfurt, 1982, pp. 83–84, emphasis mine)

What is meant by these 'richer ways' in which cares lend an element of continuity to a life could perhaps best be explained as the role that cares play in the '*narrative'* of a life – the story that an agent (or an observer) tells about it – but in this case it can hardly be suggested that this represents a response-independent property of cares. Instead, what it highlights is that the notion of cares plays an important role in the continuity that is *imposed* on a life by an observer – even by the agent herself, as observer of her own life. Our attributability judgments involve telling a story about who an agent '*really*' is, and that story (as already examined in the context of the manipulation objection and our psychological essentialist tendencies in the previous section) necessarily imposes a kind of continuity on an agent's life. Instead of arguing that this kind of continuity is created by an underlying natural kind ('cares') one could just as easily argue the reverse: not that our essentialist tendencies reflect the reality of deep selves, but that the concept of deep selves is the product of those psychological tendencies. The very concept of 'cares'– as a 'latent' truth of who an agent is, able to be drawn on whenever necessary in order to reconcile otherwise inconsistent or contradictory behaviour with our conception of the agent's continuing moral essence – might simply be the folk-psychological tool by which we, as observers, *impose* the continuity in question. Narrative continuity is not only a requirement of our moral framework (necessary in order to justify attributing responsibility for past acts to present agents) but it is also imposed on the world by a feature of our cognition recognised as psychological essentialism. To suggest that 'cares' can be identified, as objective, response-independent properties of an agent's psychology, by virtue of the role they play in creating continuity in an agent's life arguably gets things backwards.

Frankfurt's next suggestion is that volitional necessity derived from cares (as opposed to those volitional incapacities caused by other, 'external' factors like compulsion or addiction) is in some important sense "*liberating rather than coercive – i.e. […] it supports the person's autonomy rather than being opposed to or independent of his will*" (Frankfurt, 1982, p. 88). The proposed identifying feature, then, of cares-based volitional necessity, is a subjective

experience "*of liberation and of enhancement*" – explained by Frankfurt by the fact that an agent's relationship towards a cared-for object (or ideal, etc.) "*tends towards selflessness. His attention is not merely concentrated upon the object; it is somehow fixed or seized by the object. The object captivates him. He is guided by its characteristics rather than primarily by his own*" (Frankfurt, 1982, p. 89).[48] But this sense of enhanced autonomy that comes with acting on cares is also not a response-independent criterion by which we might identify the relevant deep mental states: whether or not we find an agent's behaviour (or our own) to be 'liberating' or 'autonomy enhancing' in the relevant sense will depend on our own (normative) judgment as to whether the object in question is *worth caring about*. If we do make such a normative judgment, we will quite naturally perceive acting in accordance with the corresponding desires as acting in an 'authentic', 'autonomy-enhancing' way.

The result is that whether or not we perceive a certain case of volitional incapacity in this way is ultimately determined by (and therefore tracking) our own *conscious* normative judgments as observers.[49] We have no difficulty imagining Martin Luther's actions as autonomy-enhancing and therefore based on cares, but consider just how implausible we find the case of a truly 'willing' addict – one who not only consciously endorses his actions, but does so *wholeheartedly*. This difficulty, I suggest, is due to our underlying normative judgments about how much each of those courses of action is worth caring about – a judgment, if you will, concerning the moral 'affordances' of the situation. In this case, because we judge the object of Luther's attitude to be worth caring about, we are inclined to attribute the volitional incapacity in question to his deep self. With the addict on the other hand, where we judge the object in question to be of no value, we reach for an explanation that does not require us to additionally explain why the agent in question cares about something that is on its face not worth caring about: the volitional capacity is therefore not attributed to his deep self at all. If, alternatively, the supposedly 'response-independent' identifier of cares is the agent's *own* experience of liberation or enhancement, it seems clear that we will simply be tracking what *they* consciously

[48] See (Varga, 2011) for a more in-depth explanation of the link between 'caring' about something and the experience of moral behaviour or selflessness.

[49] What exactly this sensation amounts to is not entirely clear on Frankfurt's account, in which a variety of vague descriptors are used, but a reasonable explanation might be that it is a result of the value we place on 'living in accordance with the facts' – a term borrowed from (Wolf, 1981) that is used to describe one potential reason, according to Strawson, why we might be tempted to give up our reactive attitudes in the face of the truth of determinism (see also (Wallace, 1994)). Importantly, this is a feature of our *conscious, evaluative judgments* concerning whether our chosen way of living/acting is an appropriate response to (or even 'justified by') objective facts about the world, and so even the unwilling-but-deluded addict (who consciously identifies with his actions but does not do so, as Frankfurt would say, 'wholeheartedly') is likely to experience it.

endorse, in a way that deep self theorists themselves acknowledge as flawed for all of the various reasons raised in the previous sections.

A more recent proposal relies on a distinction between ordinary desires (e.g. 'the desire for *x*') and *Y*-desires (e.g. 'the desire to quell the desire for *x*') such that cares-based volitional incapacity can be distinguished from other forms of agency-negating volitional incapacity because the latter cases will always manifest themselves in the form of *Y*-desires (Shoemaker, 2003). According to Shoemaker the unwilling addict would, if given the opportunity, take a magical pill to eliminate his desire for the drug. An agent who truly cared about something, on the other hand, would want to *continue* to care about it – what has since been described as the 'commitmental' effect of caring (Sripada, 2016). But where does the desire to *keep caring* about something come from? Quite clearly it comes from the judgment that the cared-for object is itself worth caring about – as just discussed in the context of Frankfurt's claim about the experience of being guided by the qualities of the object.[50] In this case though, any desire that I consciously endorse will have the corresponding commitmental effects simply because I have *consciously* committed myself to the *value* of the desired object as independent of my desire for it.

Consider once again the unwilling-but-deluded addict (who consciously endorses his actions, but not wholeheartedly) who normative evaluation of the world tells him that his actions are, all things considered, the thing to do. If asked whether or not he would take the hypothetical pill to magically relieve him of his desire for drugs, he would respond that he would rather continue down the path that he is on.[51] So the proposed criterion once more provides us with nothing more than a way of tracking what desires an agent consciously endorses – which, as already outlined, does not provide for a satisfactory account of our patterns of attribution. If, on the other hand, we are to rely not on the agent's self-report but instead

[50] One might attempt to argue that an agent might desire to keep caring for something for purely instrumental reasons (i.e. based on the belief that the agent will be better off if she keeps caring for it, but recognizes it as having no intrinsic value) but the cares-based theorist's response would simply be that in such cases that agent does not *really* care about the object in question at all.

[51] I readily acknowledge that these scenarios are somewhat implausible, but that is precisely the point: we find it extremely difficult to imagine that the addict *could* refuse such a pill because we find it so difficult to believe that an agent could consciously endorse the addict's lifestyle – again reflecting our own normative evaluation of the situation. The addict's lifestyle simply does not present the kind of normative 'evaluative affordances' (for us) to believe that an agent could judge it best, but that doesn't mean that there aren't people who think of themselves as 'willing addicts'. Consider, for a less normatively charged example, the overweight gourmand: just how plausible do you find it that he would refuse the desire-eliminating pill? The answer will likely depend on whether or not you believe that all fat people must secretly resent their bodies and struggle with their desire for food, but regardless of your own normative/evaluative standpoint, there will be agents who consciously endorse the opposite.

insist that there is a theoretical 'fact of the matter' as to what an agent, if given the opportunity to simply eliminate certain desires, would do, then we fall into the opposite problem: there can be no objective assessment of the counterfactual in question, and so we are forced to rely on our own judgment of what an agent *would* do, which is inevitably skewed by our own normative assessment of the situation.

I turn finally to two different accounts that attempt to provide a response-independent characterisation of the morally relevant subset of desires in terms of their 'intrinsic' nature: Sripada's 'conative self-expression' account (Sripada, 2016) and Arpaly and Schroeder's 'spare conativism' (Arpaly & Schroeder, 2013).[52] The two very different accounts are particularly interesting both because they are highly committed to the description of the deep self in terms of desires alone[53], and because they are most explicitly committed to what I have thus far described as the realist 'assumption'.[54] Both are also initially promising in their recognition of the shortcomings of many existing deep self accounts of moral responsibility – which they characterise as excessively 'rationalistic' for the reliance placed on agents' conscious evaluative endorsement/self-reflection. Ultimately though both cases for an 'intrinsic desire' view of morally responsible agency

The starting point for both accounts is the recognition that existing desire-based accounts of the deep self involve a range of 'concessions' – meaning that they are not entirely committed to the primacy of desires in constituting agents' deep selves. These concessions represent the various ways in which previous theorists have attempted to work around the problem first identified by Frankfurt, and consider desires to be relevant only, for example:

> When nestled in a hierarchy of desires [see e.g. (Frankfurt, 1971, 1982, 1987)], or when they would be endorsed upon reflection [see e.g. (Tiberius, 2002)], or held under full information [see e.g. (Railton, 2003)], or when they are derived from beliefs about what it would be rational to desire [see e.g. (Smith, 1994)], or are otherwise limited, managed, contained [see e.g. (Brandt, 1979; Hubin, 2001, 2003)]. (Arpaly & Schroeder, 2013, p. 5 endnotes incorporated)

Part of what is problematic about these various 'concessions' is that they allow the concept of a desires-based deep self to be used flexibly, so as to rationalise our moral

[52] Strictly speaking Arpaly and Schroeder do not consider their view of morally responsible agency to be 'cares-based', but they argue instead that cares-based talk can ultimately be reduced to what they describe as 'intrinsic desires'.

[53] See e.g. Arpaly and Schroeder's case for "an unabashedly desire-centred moral psychology" (Arpaly & Schroeder, 2013).

[54] See e.g. Sripada's description of his 'actualist' notion of the deep self (Sripada, 2017a), and Arpaly and Schroeder's discussion of intrinsic desires as a 'natural kind' (Arpaly & Schroeder, 2013, pp. 143–146).

responsibility attributions in particular cases, similarly to the range of desires-based views considered above. Sripada's response to this observation is his 'conative self-expression' account of morally responsible agency. His account uses the language of 'cares' to describe the subset of desires that are definitive of morally responsible agency – such that an agent is responsible for an act or attitude to the extent that it represents one of her cares – and cashes this out in terms of 'intrinsic desires'. To desire something intrinsically, in Sripada's terms, is essentially to desire it for its non-relational properties – meaning to desire it for its own sake. In this sense, intrinsic desires can be thought of as existing at the top of (and structuring) a kind of 'motivational hierarchy'. This he explains with the example of Katya:

> [Katya] wants to get on the bus. She does this because she wants to get to class, and this too is done in the service of a number of further desires: fulfilling the organic chemistry requirement, getting her medical degree, becoming a competent physician. When we trace sequences such as this to where they lead, we encounter at their very root a distinct class of conative states: cares. Katya wants to become a competent doctor because she cares about helping those who are in need – she wants to relieve their suffering. (Sripada, 2016, p. 1208)

But this formulation of 'intrinsic desire' is problematic for two reasons: an epistemic reason, and an underlying substantive reason. The epistemic problem is reasonably easily formulated in terms of the above example: how do we know, in a case like Katya's, where the instrumental hierarchy of desires stops? Were we, as observers, to be slightly more cynically inclined, we might wonder whether the desire to relieve the suffering of others was really at the foundation of Katya's motivational hierarchy (or whether it was the intrinsic desire structuring her reward system). A Nietzsche, for example, might suggest that this is merely a presentable façade to the underlying will to power that drives Katya; a Freud might suggest that it all reveals an underlying need to win the approval of her doctor father. The point is not that one of these is the answer but that any of them could be, and Sripada's account doesn't provide us with any *a priori* criterion for identifying the correct one.

In light of this epistemic problem, it seems likely that our attribution of 'intrinsic desires' will be influenced by our own judgments as to what typically does motivate people – which will in many cases coincide with normative judgments. If we have to guess (which we inevitably do) what 'intrinsically' motivates an agent, in the absence of any definitive evidence from the agent's behaviour, we are most likely to provide an explanation in terms of what we (consciously) judge to be intrinsically motivating. What this turns out to be will likely be influenced by a normative understanding of human behaviour – drawing on both salient prescriptive (moral) and descriptive (statistical) norms. Our moral frameworks provide us with

one obvious source of explanations, where what is 'good' in a moral sense is understood to be intrinsically motivating, and in many cases will readily provide a suitable explanation. In cases that cannot be made to fit the pattern of 'moral' behaviour, a descriptive understanding of other aspects of 'human nature' provides explanatory alternatives: we understand that agents may act on reasons such as selfishness or greed, and so can still attribute behaviour that fits such a pattern to a kind of 'intrinsic motivation'.[55]

The underlying substantive problem is potentially more serious: why should we assume that all of an agent's desires are organised in a hierarchy, with particular 'intrinsic' desires providing the structure? Certainly this fits with our folk-psychological rationalisations, where instrumental reasons feature prominently in our conscious explanations. It even seems plausible that, if asked enough times 'why do you want *x*?' an agent will eventually no longer be able to provide further instrumental explanations. This seems to be the case with Katya, who no doubt cannot explain why it is that she desires to relieve others' suffering – without, that is, referring to the purported properties of the object of that desire, i.e. the fact that it is *good* to do so. But what moral significance (indeed what real psychological significance) can there be to the point at which we cease to be able to provide answers in terms of relational properties (within our own psychology) and are forced to refer to postulated intrinsic properties of the object of some desire? Certainly all the Freudian or Nietzschean analyst seems to be doing is pushing the explanation in terms of relational properties a little further. And where the 'reasons' explanations typically consciously available to agents run out, one can reliably turn to causal-historical explanations for the existence of particular desires.

In a sourcehood-compatibilist moral framework, the existence of any desire must always be explicable by reference to some antecedent conditions – some complex combination of 'biological' and 'environmental' factors – such that the idea of an 'intrinsic' desire becomes quite incomprehensible. This is not to say that the sourcehood incompatibilist is not entitled to invoke intrinsic desires. However, once it is acknowledged that 'intrinsic' in this context cannot mean that a desire has its *ultimate* source within an agent (this is the basic source-incompatibilist argument) but instead only describes how causal pathways may operate *through* an agent in an appropriate way, then the source-incompatibilist cannot invoke the idea of 'intrinsic' desires to explanatory work without putting forward a positive account of what

[55] In many cases there will be significant overlap between moral and statistical norms, as our beliefs about 'normality' are commonly a composite of the two – see (Bear & Knobe, 2016).

'intrinsic' in this context might mean. Sripada, however, is attempting to justify his account of the deep self by reference to 'intrinsic' desires without providing such a positive account. In light of the epistemic difficulties outlined above, Sripada is essentially asking us to accept that there are intrinsic desires – in a sense relevant to morally responsible agency – because sometimes our folk-psychological instrumental explanations run out, without providing any link between the limits of our folk-psychological explanations and morally responsible agency. The causal-historical explanation of the existence of any particular desire is necessarily an explanation in terms of the relational properties of the object of the desire. To suggest that something is desired 'intrinsically' because we no longer have any folk-psychological reasons-based explanation for its existence is to attribute far too much significance to the limits of our conscious introspection and folk-psychology. It ultimately runs into the same difficulties as the various accounts criticised by Sripada for being too 'rationalistic': the bounds of morally responsible agency appear to be set according to the limits of an agent's conscious, self-reflective activity (or indeed by the limits of our folk-psychological imagination).

To illustrate, consider an earlier example of 'acculturated attitudes': an agent brought up in a puritanical environment has particular attitudes towards pre-marital sex such that she desires to wait until she is married, and *resist* her sexual desires in the meantime. If asked *why* she has this particular desire, she might answer that she desires it because it is right or good. Alternatively, she might answer that she desires to wait until she is married because she desires to please God, and that she desires to please God because to do so is right and good. Regardless, the sequence will most likely end with some explanation in terms of non-relational properties of the desired object – with an 'intrinsic desire' for which no further explanation is forthcoming.[56] Whilst we know that the non-puritanical observer is likely to conclude that what the agent *really* wants is represented by her sexual desires, not by the (acculturated) desire to repress them (Watson, 1975) we can equally well imagine that to other members of her community the kind of 'intrinsic-desire' analysis suggested by Sripada would yield the opposite result. The idea of an 'intrinsic desire' is in such cases quite clearly a manifestation of the limits of the relevant folk-psychological imagination, as opposed to anything strictly psychological.

[56] Note that these final 'intrinsic desire' explanations need not be 'moral' as in the example above. I might desire to go to a particular restaurant, because I desire to eat a particular dish, because that particular dish *tastes good*. Here too it would be possible to provide a causal-historical explanation as to why I desire to eat that particular dish – some combination of evolutionary and cultural factors that explain why the particular combination of textures and flavours and smells appeals to me – but this is undoubtedly *not* the kind of explanation I would be likely to provide.

Arpaly and Schroeder's 'intrinsic desire' account of the deep self employs a very different (and slightly unusual) notion of 'intrinsic'. Within their 'spare conativisim' an intrinsic desire is simply a 'state of the reward system' – meaning, essentially, something that an agent's psychology 'constitutes as a reward'. Intrinsic desires are understood to be the natural kind underpinning all of the motivational, emotional and cognitive effects understood in psychological terms to be the products of the reward system: if attaining *x* feels rewarding, this is because of some intrinsic desire for *x*; if the thought of attaining *x* feels motivating, this is similarly due to an intrinsic desire. All we are doing in terms of our moral responsibility attributions, according to Arpaly and Schroeder, is trying to identify the intrinsic desire that rationalises a particular behaviour – trying to separate out the intrinsic from the 'instrumental' or 'realiser' desires – so that we may attribute it to the agent. To have 'good will' on this account is to have an intrinsic desire for 'the good' (however it is to be characterised) in the sense that one's reward system constitutes 'the good' as a reward.[57]

It is worth noting first of all that such an account suffers from an epistemic problem similar to that faced by Sripada's account above: it will always be underdetermined, from an agent's behaviour, exactly which 'intrinsic desire' rationalised the behaviour in question or indeed how the object of that intrinsic desire is 'conceptualised' by the agent's reward system. It is likely, in the face of such uncertainty, that our judgments concerning an agent's intrinsic desires will be similarly influenced by our normative judgments and available folk-psychological explanations.

The epistemic issue aside though, Arpaly and Schroeder's account is in some ways a 'bullet-biting' approach to a psychologically descriptive account of the deep self. By reducing the deep self to the language of psychology, however, the account proves extremely difficult to reconcile with our practices of moral responsibility attributions. Take, for example, some quite standard cases of non-attributable desires: pathological compulsions like those involved in Tourette's Syndrome, and addictive behaviour. In the former case, Arpaly and Schroeder argue that the compulsive behaviour in question simply doesn't involve any 'action' by an agent – because it is not a product of the reward system and so not 'rationalised' by reference to any

[57] Arpaly and Schroeder make the additional claim that 'good will' further consists in conceptualizing the good in the right way – that is to say, that 'the good' itself constitutes a reward in the agent's reward system, rather than things that happen to be good simply independently constituting rewards for an agent. Not only does this impose a 'rationalistic' framework on a supposedly purely desires-based conception of moral agency, but it is also simply quite difficult to imagine, when desires are reduced to simple 'states of the reward system' (such that, to simplify Arpaly and Schroeder's account somewhat, something is desired if its attainment or anticipation produces dopamine) how the object of the desire is 'conceptualised' at all.

state of it – but instead mere 'movement'. Not only does this raise the concern that the purely 'desires-based' account of morally responsible agency is introducing some 'concessions' (i.e. the desires relevant to morally responsible agency must be 'reasons-responsive') but it is unclear on what principled (objective, response-independent) basis the distinction between 'action' and 'movement' is being drawn. In the case of addiction, Arpaly and Schroeder offer a complex explanation of how addictive substances 'hijack' the reward system, based on the claim that the 'reward' produced by drugs (i) operates without any cognitive/perceptual mediation, bypassing the role of expectations in 'computing the reward signal' and (ii) always creates a learning reaction "*as though more strongly desired things happened than were unconsciously predicted*". As to the first claim, the example used to distinguish the effects of 'ordinary stimuli' on the reward system from those of addictive drugs is the following:

> A person who hates Imelda will not have dopamine release promoted in her by seeing Imelda. But a person who hates cocaine in just the same way, with just the same neural connections between the idea of cocaine and the reward system as her counterpart has between the idea of Imelda and the reward system, will nonetheless get a strong surge of dopamine-type effects if that person consumes cocaine. This is the way in which addictive drugs hijack the reward system." (Arpaly & Schroeder, 2013, p. 278)

The idea that drugs operate in an entirely different way to any other kind of stimuli does not necessarily seem borne out by this example. First of all, it does not seem impossible that with respect to 'normal stimuli' I might claim to hate something, and indeed experience feelings of dread in anticipation of that thing, and nonetheless enjoy it in the end. This might be a way that many people feel about certain kinds of social outings. At the same time, if I 'hate' something but nonetheless get a strong surge of dopamine-type effects every time I experience that thing, people might reasonably start to doubt that I actually hate it. How far would Arpaly and Schroeder be willing to stretch the addiction-style justification? Would it cover any case where the *idea* of something disgusts somebody but the experience produces 'dopamine-type' effects? Would it extend to the hatred a paedophile might feel for his actions? The hatred an obese person might feel for food? The hatred internalised homophobia might make someone feel about homosexual sex? The idea that there is a necessary parallel between the neural connections linking the *idea* of something to the reward system and those linking the actual thing to the reward system, broken only in the extraordinary case of addictive drugs, is clearly implausible, and to suggest that it represents a morally relevant distinction between attributable and non-attributable agency goes counter to our responsibility intuitions in the remaining cases. The second alleged 'special feature' of addictive drugs rests on a very doubtful empirical claim

– indeed the way in which regular drug users often find themselves forced to take greater and greater doses to achieve the same 'high' suggests the exact opposite – both that the prediction mechanisms do have a role to play, and that sometimes the result of drug use can be something akin to 'disappointment'.

Overall, what Arpaly and Schroeder's account shows us is the trade-off involved in a bullet-biting approach that attempts to describe the contents of the deep self in purely psychological terms: it becomes extremely difficult to fit the resulting theory to our attributability intuitions, such that recourse needs to be had to case-by-case explanations based on speculative neuroscientific claims. Indeed the authors recognise, in their conclusion, the wide range of 'difficult cases' that a theory of morally responsible agency has to cope with, and can only suggest that such cases be tackled individually, 'with the help of empirical science'. It is unclear, however, whether or not even the cases of non-attributable desires already discussed have been successfully explained by empirical science. One alternative interpretation is that the language of empirical science has simply been co-opted to explain away our more complex intuitions.

Ultimately the various accounts I have considered within the Frankfurtian tradition of deep self moral responsibility have collectively failed to come up with any response-independent identifier of a particular set of mental states that would explain our patterns of responsibility attribution and justify the realist assumption. A common feature of the various proposals for differentiating 'deep' mental states – 'cares' – from mere desires was that they either required us to rely on the self-report of the agents in question (in which case they were ultimately influenced by the agents' reflective self evaluation and therefore mere conscious acts of identification/endorsement) or they required us to rely on our own assessment of an agent. In the latter cases it is all too clear how our various cognitive biases (involving the construction of continuity and the projection of normative judgments) are at work in creating the very features of the deep self that are used to justify our attributability intuitions.

I have also argued that fully accepting the consequences of the potential truth of determinism, as must be done for any compatibilist theory of moral responsibility, makes it impossible to conceive of the line between 'internality' and 'externality' posited by sourcehood compatibilists being anything other than an artificial one, imposed on the world by us as observers. This is not to deny that some causal pathways operate through agents' mental states whilst others do not – so much is trivially true, and there is a straightforward division between the mental life of an agent and other kinds of events in the world that nobody would deny. The

kind of internal/external distinction posited by sourcehood compatibilists, however, attempts to distinguish between two different ways in which causal chains may operate through an agent's mental states: in the one case giving rise to deep (morally responsible) agency and in the other case simply passing through superficially. It is my contention that the deep self theorists considered so far have failed to provide any robust criterion for making such a distinction. In the next section I will apply a similar analysis to the Watsonian tradition, arguing that the flaws examined here are mirrored there, before turning to consider just how deep self theorists have managed to leave unchallenged the by now obviously problematic realist assumption.

§3) The Watsonian Approach

Watson's initial approach to the deep self and moral responsibility begins, like Frankfurt's, with a rejection of the incompatibilist thesis captured by the 'principle of alternate possibilities' – though his approach is in some way the opposite of that taken by Frankfurt. Whereas Frankfurt examines cases in which an agent is *not* free to act otherwise than she in fact does but her actions are still attributable to her, Watson begins with the observation that there are many instances where agents appear to *be* free, in the incompatibilist sense, to do otherwise and yet we do not attribute their actions to them. His objection to the compatibilist conception of the kind of 'free' action required for morally responsible agency is that it would end up conflating free action with intentional action: any intentional action is necessarily an action that an agent is *able* to do, and thus that she is free to do on this account, in a way that leaves no room for explaining how *some* intentional actions – those resulting, for example, from addictions, manias or phobias – may be, in an important sense, *unfree* (meaning not attributable to the agents).

The alternative, compatibilist account of free action and morally responsible agency introduced by Watson involves drawing a distinction between what an agent *wants* and what she *values*, where 'valuing' refers to what an agent *most* wants or *truly* wants in a sense that his account aims to elucidate. To begin with though, Watson must overcome an apparent difficulty: as he is quick to point out, the desire by which an agent is moved to action is, in a very obvious sense, her strongest desire, or what she wants *most*. So his account must come up with a sense of what an agent 'most wants' that is distinct from the mere relative strengths of desires. This work is done by the concept of 'valuing' and the particular Platonic conception of practical

reasoning in which it is located.[58] Watson distinguishes between an agent's *valuational* and *motivational* systems: the motivational system describes all those 'considerations' (psychological elements) that incline an agent towards action in varying degrees; the valuational system describes "*that set of considerations which, when combined with [an agent's] factual beliefs (and probability estimates), yields judgments of the form: the thing for me to do in these circumstances, all things considered, is x*" (Watson, 1975). The possibility of free (morally responsible) agency exists only in beings with such a valuational system, and the corresponding possibility of unfree agency arises from the potential for inconsistency between the two systems – that is, where there are factors within an agents motivational system that move her to action *despite* the judgments of her valuational system, i.e. contrary to her all-things-considered judgments about what is best in a given situation.

The deep self, on Watson's model, is constituted by the valuational system. Judgments of the valuational system produce corresponding desires in the motivational system which, in the case of free agency, will be the desires that end up moving the agent to action. When we attribute responsibility for an action to an agent, it is because we judge that action to reflect what she *values* – i.e. to ultimately stem from a judgment of her valuational system. In cases where we attribute responsibility for a 'bad' action, this may either be because we judge the agent to have valued the wrong thing (vicious action) or not to have valued the right thing enough (weak action). It is in this way, on Watson's account, that our (correct) deep self assessments underpin (explain and justify) our patterns of moral responsibility attribution: they identify the facts about an agent's valuational system.

The first difficulty that Watson must come to terms with is the problem of self-deception, often presented as the result of 'acculturation'.[59] He considers the example of an agent raised in a puritanical context who holds particular conscious attitudes towards pre-marital sex but

[58] Watson contrasts his 'Platonic' conception of practical reasoning with a 'Humean' conception. As we will see later, the terminology of 'Humean' and 'Platonic' is adopted by Shoemaker when describing the two competing traditions in the deep self debate (Shoemaker, 2015a) which I have labelled Frankfurtian and Watsonian. Watson, however, is not using the term to refer to the alleged 'substance' of the deep self (desires/cares on the one hand as opposed to evaluative judgments/commitments on the other) but to the structure of the agential psychology in which such elements are located. On a Humean account of practical reasoning there is a single source of motivation – *desires* – and the faculty of reason plays a purely instrumental role in the calculation of how best to satisfy those desires. On a Platonic account there are two (competing) sources of motivation, with the faculty of reason (forming judgments about the good) constituting an independent source of motivation from an agent's mere appetitive desires. As Watson observes, it is only in the latter case that the possibility of free or unfree action arises.

[59] The idea of 'acculturated' attitudes as distinct or distinguishable from the 'deep' attitudes that are essential to autonomy or morally responsible agency has had a relatively broad influence – see, for example, (Meyers, 1987).

nonetheless *feels* conflicted on the matter. We might ultimately consider that her conscious attitudes do *not* reflect her 'true' evaluative judgments, which we might think are manifesting themselves in desires that conflict with her 'acculturated' attitudes. As Watson says:

> Acculturated attitudes may seem more akin to evaluation than to appetite in that they are often expressed in evaluative language ("divorce is wicked") and result in feelings of guilt when one's actions are not in conformity with them. But, since conflict is possible here, to want something as a result of acculturation is not thereby to value it, in the sense of 'to value' that we want to capture. (Watson, 1975, p. 215)

Already then we can see a shift *away* from the idea of consciously held attitudes being relevant to the deep self. Whilst the Frankfurtian characterisation of the deep self in terms of desires has been replaced with 'evaluative judgments', the need to provide some kind of qualification to that class – to identify a particular *subset* of evaluative judgments relevant to morally responsible agency – remains. Watson's initial qualification – restricting the relevant evaluative judgments to those made "*in a cool and non-self-deceptive moment*" (Watson, 1975, p. 215) – is unfortunately clearly inadequate. Not only does the 'cool' qualifier run counter to a number of intuitions (concerning cases where an individual's 'true' self appears to be revealed in the heat of a particular situation, see (Sripada, 2016)) but the qualifier of 'non-self-deceptive' is obviously circular: it simply proposes that the relevant subset of evaluative attitudes that are attributable to an agent's deep self are those that in fact accurately reflect that agent's deep self. The question is then whether there is any other, response-independent, non-circular feature capable of identifying the relevant 'deep' evaluative judgments. If not, then we must ask what, if anything, is achieved through the reframing of the deep self in terms of evaluative judgments (as opposed to desires) in the first place.

(i) Attributability in Passive Cases

As the acculturation objection demonstrates, there are some clear cases where an agent's 'true' evaluative judgments (in the sense relevant for morally responsible agency) are at odds with her conscious (evaluative) attitudes – that is, we do not attribute those conscious attitudes to her deep self. Similarly, there are situations in which we find it natural to attribute responsibility to agents for behaviours that don't seem to be the product of evaluation at all: what an agent notices or neglects, spontaneous attitudes or whims, forgettings or omissions, involuntary reactions, etc. These instances of 'passivity' bear a strong resemblance to the kind of volitional necessity or incapacity discussed by Frankfurt – cases where, despite what an agent has consciously endorsed, she may in a sense surprise herself by what she ends up doing in a

way that doesn't seem to fit the description of compulsive action. Consider the case of Huck Finn: when faced with the prospect of handing Jim over to the slave hunters, though he clearly thinks this to be the best (most moral) course of action, he finds himself unable to do it. He feels that he is 'passive' in the face of some apparently psychologically constraining force, and yet we clearly find his actions to reflect something important about his deep self.[60]

The range of cases of attributable passive behaviour is much wider than those Huck-Finn style cases involving obvious inner conflict between evaluative (bad) 'morality' and unarticulated 'sympathy'. We may, at least potentially, take cases such as an agent's forgetting a friend's birthday to tell us something about who she is. As Angela Smith observes:

> We often respond to people's spontaneous attitudes, reactions and unreflective patterns of awareness in many of the same ways that we respond to their voluntary actions. […] [M]any of the features upon which our own judgment, or the judgment of others, may pronounce fall outside the scope of our immediate voluntary control. [Involuntary responses] provide an important indication of a person's underlying moral commitments, of who he is, morally speaking. Their alleged passivity, far from undermining their attributability to persons, may actually be the strongest mark of their genuineness and sincerity. (Smith, 2005, pp. 241–242)

A partner's (passive) failure to notice a loved one's preferences, for example, may be both blameworthy and specially attributable to the agent in question. But how do we account for this in terms of 'evaluative judgments'?[61]

Smith, like Watson, defends a 'rationalist' (Watsonian) criterion for moral attributability, according to which an agent's behaviour must reflect her evaluative judgments in order to be attributable to her. But the agents in these situations clearly take themselves to be committed to something very different to what is expressed in their behaviour. Smith's explanation is that there is simply a normal and necessary connection – a rational relationship – between what we value and what does or doesn't occur to us or prompts spontaneous reactions in us. As such, when we fail to notice a loved one's preferences, or forget an important birthday or anniversary,

[60] See (Bennett, 1974) for early discussion of the case, and for further development, including the explanation of how the responses are neither directly nor indirectly under an agent's control, see (Adams, 1985). Though this particular case involves a degree of reflection (making it slightly different to cases of omissions, etc.) the central feature of 'passive' cases is that, from the perspective of the agent, it is done 'despite himself' – which can mean either in the absence of conscious, intentional decision-making, or indeed despite or contrary to the apparent outcome of such conscious, reflective activity.

[61] Smith notes that her use of the term 'evaluative judgment' is somewhat loose, often replacing it with 'evaluative commitment'. What this aims to reflect is that fact that she is explicitly *not* talking about consciously held propositional beliefs, but about "tendencies to regard certain things as having evaluative significance" such that "although I may never have consciously entertained these evaluative judgments, I see that they are correctly attributable to me in virtue of my own responses to the situations I confront" (Smith, 2005, p. 252).

"these cases can reasonably be taken to reflect a lack of appreciation for the significance of the events in question". That is, *if* one valued the other's happiness etc. *enough*, one simply would not have had the involuntary response in question in the first place. This is all in virtue of what Smith describes as a

> Direct rational connection between our spontaneous reactions and our underlying evaluative judgments and commitments. […] To feel contempt towards some person, for example, involves the judgment that she has some features or has behaved in some way which makes her unworthy of one's respect, and to feel regret involves the judgment that something of value has been lost. There seems to be a **conceptual connection** between having these attitudes and making, or being disposed to make, certain kinds of judgments[…] Unlike brute sensations, which simply assail us, our spontaneous reactions reveal, in a direct and sometimes distressing way, the underlying evaluative commitments shaping our responses to the situations in which we find ourselves. (Smith, 2005, p. 250)

When these 'underlying evaluative judgments' are revealed to us in the potentially disturbing way described by Smith we may, of course, try and change them: we can assess our evaluative judgments, asking whether they are properly justified, and try to modify or abandon them in the absence of any such justification. This presents an interesting difficulty though: if unsuccessful in altering our spontaneous reactions, the 'conceptual connection' between evaluative judgments and those spontaneous reactions breaks down in a kind of *failure of rationality* for which Smith admits that she has no explanation;[62] if successful in altering them, it means that our 'all-things-considered' view of what is best is something *other* than that evaluative judgment that seemed initially attributable to us. In the latter cases we are likely to take the later judgment (and the modified behaviour that results from it) as reflective of an agent's 'true' evaluative commitments, and the earlier ones will be explained away as something other than a true/deep evaluative judgment. If the candidate response-independent feature of 'deep' evaluative commitments was that they express themselves in our spontaneous attitudes and behaviour, then we are at a loss to explain how the agent's *initial* spontaneous attitudes reflected something other than her all-things-considered judgment of what is best.

So when does our spontaneous behaviour *not* reflect our evaluative commitments? One potential (though by no means comprehensive) answer is when it is 'acculturated'. Consider the agent brought up in a deeply racist environment, who has never thought much about race but certainly wouldn't consider himself a racist: his spontaneous attitudes might include certain involuntary fear responses towards others; perhaps he slips into the use of racial epithets without

[62] See (Smith, 2005, p. 255).

thinking. When all this is pointed out to him he might be shocked and disturbed; he might seriously reflect on the matter and determine to eliminate all such spontaneous behaviours to reflect his 'evaluative commitment' to racial equality. But perhaps he will be unsuccessful in totally reforming his spontaneous behaviours. Perhaps implicit association tests (IAT[63]) would reveal spontaneous fear responses or other negative associations towards people of other races. Is this person a racist? Can we attribute to him the underlying evaluative judgments that his behaviour seems to reveal, even if in every aspect of his conscious decision-making he manifests a commitment to racial equality? Whilst some theorists might suggest that we can,[64] this might risk overlooking or denying our more complicated intuitions towards the case: not only can we easily imagine ways of framing spontaneous reactions as 'acculturated' (a product not of deep characteristics but of an agent's flawed surroundings) but where spontaneous reactions in something like an IAT test appear to be in conflict with an agent's everyday behaviour and conscious motivations it is not immediately obvious that the former is attributable to the agent over the latter.

Once again it would appear that manifestation in spontaneous behaviour, whilst it may appear as a rationalisation for our moral responsibility attributions, is in reality far from conclusive and certainly not a sufficient or even necessary condition for attribution. It operates instead as one of a range of available rationalisations which we may draw on to justify our attributability intuitions. A readily available alternative explanation in terms of 'acculturation' can overcome our 'general' intuition that some spontaneous attitude reveals or expresses the deep self. This might be even more likely when the case in question is particularly close to home: it has been suggested that many self-described unprejudiced people might be revealed by IAT studies to make implicit associations that on the 'rational relations' view endorsed by Smith would reveal underlying evaluative judgments that they themselves would disavow. When neither an agent's consciously held propositional beliefs nor the subconscious elements of her psychology revealed in her spontaneous behaviour are reliable guides to what her 'true' evaluative commitments are, it seems that a response-independent feature capable of identifying the deep self and supporting the realist account remains elusive.

[63] See (Greenwald, McGhee, & Schwartz, 1998).

[64] See, for example, (Brownstein, 2016).

(ii) Perverse Cases

Another set of cases from which interesting parallels may be drawn with the Frankfurtian account (and its failure to identify a response-independent feature constitutive of 'deep' attitudes) is what Watson has described as 'perverse cases' (Watson, 1987a). These are cases where an agent *consciously* judges one particular course of action to be best, and at the same time, with no apparent conflict, *consciously* decides to do something else. Watson offers the example of the philosopher who judges it best, all things considered, that he stay at his desk and write, but nonetheless decides to go golfing instead and (according to Watson) fully embraces the decision to do so. As a result, the actions involved in perverse cases are "plainly neither cases of compulsion nor weakness of will". To see what they are though, we must first work out what compulsion and weakness of will are on the Watsonian model of the self, considering Watson's 'Skepticism about Weakness of Will' (Watson, 1977).

The scepticism that Watson attempts to counter with respect to weakness of will is the following: if weakness of will consists in giving in to desires that an agent is in fact capable of resisting, how can we make sense, within his model of the self, of an (unrealised) capacity to resist? Compulsive behaviour, explained in terms of the valuational and motivational systems, involves a strong desire in the motivational system that overwhelms those desires created by the valuational system. Resisting a desire, in these terms, would mean an agent's valuational system creating a sufficiently strong corresponding desire within the motivational system that outweighs the recalcitrant desire. But the strength of an (endorsed) desire within the motivational system directly corresponds to the extent to which the object of that desire is *valued* within the valuational system. An agent cannot, by force of will, *change* how much they value something; her evaluative judgments are *constitutive* of her self. Weakness of will and compulsion are essentially identical in so far as both involve a desire within one's motivational system outweighing a competing desire produced by the valuational system. Not only can we not make sense of an unrealised capacity to resist on this model of the self, but the very idea of unrealised possibilities as a morally relevant factor should surely be impossible in a compatibilist theory of moral responsibility. In a compatibilist framework any judgment of moral responsibility must stand even if it turns out that things could never have been otherwise. Weakness of will as a moral category, on any compatibilist account, needs therefore to be explained in a way that is compatible with an agent *not actually* being able to do otherwise.

Watson's solution is an account of weakness of will as giving in "to desires which the possession of the normal degree of self-control would enable them to resist" (in contrast to

compulsive desires, which are those that a 'normal' agent would be unable to resist). But what is a 'normal degree of self-control'? It is unhelpful to imagine something like 'strength of will' in realist terms – that is, as an objective feature of an agent's psychological make-up. Watson's model of the self does not require any particular ability or capacity (or 'strength') to translate elements of the valuational system into desires within the motivational system. His model deals in terms of *rational relations*: there is a necessary relationship between valuing something (judging to be best, etc.) and desiring it (in the sense of being motivated to bring it about). If this relationship itself were subject to manipulation by some further additional element of an agent's psychology – if an agent could 'choose' how much influence her values had on her motivational system – the seat of the deep self would be displaced from the valuational system to this new 'choosing' system. Certainly this is a model of the self that might appeal to motivational externalists, but it doesn't resolve the issue. If there is an ability to 'choose' how much one's values influence one's behaviour, through an exercise of will, then the problem of self-determination appears once again. And if, alternatively, one would argue not that there was a 'choice' mediating the effect of the valuational system on the motivational system but rather some fixed 'strength of will', this would ultimately be no different than the claim that some people simply *value* things more. To avoid these issues, Watson's proposal has to be understood in terms of normative expectations concerning the kinds of desires that an agent *ought* to be capable of resisting.

Think of this within a normative framework of 'good' and 'bad' desires. The desire to have a slice of pie, or a shot of heroin, might be labelled a 'bad' desire, and the desire to work on one's manuscript, or to go out for a jog, a 'good' desire. When an addict takes a drug, one way we might explain this behaviour is in terms of compulsion: the addict really *does* value the right things (i.e. wants to live a healthy, drug-free life) but is simply overwhelmed by the strength of her physiological addiction which creates an overpowering desire for the drug. The addict is therefore blameless. When a dieter takes a slice of pie, we are more likely to turn to an explanation in terms of weakness of will: the dieter values her health *a little*, but she *should* value it more – enough to overcome the desire for a slice of pie. We don't attribute the desire for pie to the agent's deep self, but we do attribute some blame for her not valuing the good 'enough'. Both cases are different from vicious action (bad action attributable to a bad deep self) where the valuational system produces a 'bad' desire which is straightforwardly attributable to the agent.

Returning to the 'perverse cases' with which we started, how can we make sense of what is going on? Certainly it seems implausible that the desire to play golf is compulsive – indeed we know from experience that 'normal' agents can and do regularly resist the desire to go golfing. But it can't so easily be explained in terms of weakness of will either, because Watson tells us that the agent fully (and consciously) endorses his decision to play golf, even whilst he simultaneously believes that staying to work on his manuscript is the all-things-considered best thing to do. He feels none of the regret or conflict about the decision that might suggest weakness of will. Is it a case of 'spontaneous behaviour' revealing the agent's underlying evaluative judgments? Certainly the agent is not caught off guard by the behaviour, and the consciousness with which he decides to go golfing means it could hardly be described as a 'passive behaviour'. All of this leaves the realist in a difficult position: there *must* be a fact of the matter about the agent's deep self – about what she values most – and yet there is no clear criterion for determining what it is. Watson himself is forced to accept that the perverse cases leave us "*with a rather elusive notion of identification and thereby an elusive notion of self-determination*" (Watson, 1987a, pp. 150–151).

Elusive as the notion of self-determination remains, the perverse cases are particularly interesting for the way in which an agent appears to simultaneously consider one course of action to be best and fully endorse the decision to take another. What is different about the way in which the philosopher considers working on his manuscript to be the best thing to do and the way in which he consciously endorses the decision to go golfing? The answer, it seems to me, reveals something interesting about the influence of an interpersonal framework of normative expectations on our assessment of the supposedly 'agent-centred' features of the situation.[65] The fact that we have so much difficulty determining the content of the agent's deep self in the perverse cases shows something interesting about the role of normative expectations in our attributability judgments. Why isn't the case simply one of the agent's 'true' evaluative commitments being revealed to him by the course of action that he ends up taking? The reason

[65] The normative expectations that I am referring to are not to be narrowly construed in terms of e.g. strictly *moral* duties and obligations between agents in a deontological framework. The wide context in which we talk about cases of compulsion, weakness of will, self-control or virtuous action often fall outside such a narrow conception of normative demands. Obviously Watson's philosopher is not under any strictly moral obligation to work on his manuscript, and yet there is still a sense in which the 'work' in question is the subject of normative expectations – those normative expectations from which we derive a common understanding of conscientiousness as virtue, hard work as elevating, etc. The framework of normative expectations does not even necessarily require a unified conception of the good. It can allow for individual ideas such that, for example, even if I am not committed to the same ideals as the philosopher (if I cannot myself conceive of the value of devoting one's life to a manuscript) I can nonetheless make moral/normative judgments attributing something like weakness of will if she fails to pursue that individual ideal.

is that it is hard for us as observers to eliminate our *own* normative judgments from the picture. *We* hold the (conscious, propositional) belief that working is the thing to do and that skipping a day of work to go and play golf involves some kind of weak-willed capitulation to mere (non-evaluative) desire. But everything in the agent's behaviour suggests the opposite: he clearly embraces the decision and shows no signs of the conflict or regret that we would expect in a case of weak-willed action. To accept that the agent's 'evaluative commitment' is really to golfing though would contradict our own normative assessment of the situation: if work *is* the thing to do, how can the agent's conscious, evaluative commitment to that idea be *wrong*, and be *revealed* to him to be wrong in the moment in which he decides to go golfing? Moreover, the agent seems to agree with us because he continues to hold the same conscious beliefs as before. In this case it might even be the agent's conscious, evaluative attitudes that he is unable to change, despite clearly endorsing (through his actions) something else.

Ultimately the perverse cases can only take us so far on this particular point. The possible conclusion, of which it gives us only an inkling, is that the notion of evaluative judgments carries with it an implication of accurately assessing (evaluating) the world, and thus of our own normative commitments reflecting an objective truth. As such, when we go to attribute evaluative judgments to others, we turn primarily to what we consider to be such objective (moral) truths, for which our shared normative framework provides the most salient examples. In the absence of any properly response-independent identifier of the agent's 'true' evaluative commitments – or of 'deep' attitudes under any other description – we have good reason to explore the possibility that our attributability judgments might be tracking, or at least influenced by, something else. Exactly what else, and just how it influences our attributability judgments, will become clear from the empirical results discussed in Chapter III.

(iii) Reconciling Deep Self Traditions

The evaluative judgments-based approach put forward by Watson was intended to be an alternative to a Frankfurtian (desires-based) approach that faced serious difficulties. Frankfurt's initial account in terms of higher-order desires seemed problematic because too many counterexamples presented themselves: it was possible to find cases where someone's 'higher-order desires' seemed to conflict with our attributability intuitions – specifically, when we could see those higher-order desires as the outcomes of some other causal force (the manipulation objection). Watson's initial account however faced similar problems: it was all too easy to find cases where we attributed attitudes to an agent that ran counter to her conscious evaluative

judgments, simultaneously attributing 'deep' evaluative judgments that she would not consciously have endorsed. In each case the original account has been modified so as to cope with a range of recalcitrant cases, with the unfortunate result that the originally tightly unified descriptive criteria for the deep self have ben replaced by increasingly vague and diffuse criteria that are flexible enough to cover all of our attributability intuitions. In this way the realist deep self accounts in question have come to provide an accurate representation of our folk-psychology – our folk rationalisations – of responsibility attribution. They reflect the realist assumptions of our folk theory, and suggest, in a reassuring way, a justification of our folk practices. It just turns out that there is no response-independent feature constituting the deep self – no natural kind capable of sustaining the realist assumption. Instead, the deep self theory ultimately arrived at by both traditions, despite their apparent disagreements, is one and the same: an unidentifiable psychological entity is postulated – drawn from our folk-psychology – which serves as a placeholder in the various rationalisations of our moral responsibility intuitions.

The picture we are left with on the Watsonian account is of 'true' or 'deep' evaluative commitments that do not necessarily coincide with an agent's conscious evaluative judgments, are not necessarily responsive to an agent's reflective evaluation, and manifest themselves in behaviour that cannot be objectively distinguished from compulsive (non-attributable) behaviour. In this sense, not only does the resulting concept of deep evaluative commitments bear very little resemblance to the idea of 'evaluative judgments' with which Watson began, but it is strikingly similar to the notion of 'cares' arising out of the Frankfurtian tradition. But why do we have two *different* placeholders – cares and evaluative commitments? Why are there two apparently competing traditions converging on the same account?

An initial hypothesis is that the evaluative judgment/desire distinction might reflect the familiar belief/desire distinction from philosophy of mind. If beliefs and desires are the basic currency of mental states, it might make sense that a tradition has developed around each in moral psychology. This hypothesis breaks down though when we consider the belief/desire distinction more closely. The basic distinction is made according to the 'direction of fit' of each kind of mental state with the world: we aim to have our beliefs conform to the state of the world, but we aim to have the world conform to our desires. Evaluative judgments and desires, in their most basic form, appear to conform to that dichotomy, with evaluative judgments representing a kind of belief about what is best (which might be subject to change if out of conformity with 'normative reality'). But the kind of desires and evaluative judgments that were the subject of

Frankfurt and Watson's initial accounts respectively do not: 'higher order desires' were desires formed on reflection about what one most wanted to be moved by, and in that sense aimed to accurately reflect a kind of inner reality; evaluative judgments were an intrinsic source of motivation. In this sense both categories fell somewhere in the middle of the traditional belief/desire distinction, and the two traditions arising out of them are not so easily explained in those terms.

Despite the ability of both 'cares' and 'evaluative commitments' to explain the full range of attributability intuitions, there remains a sense in which some cases seem to be more intuitively explained in terms of one rather than the other. This may explain the appeal behind what has been labelled an 'ecumenical approach' (Shoemaker, 2015) that attributes behaviours to an individual when it expresses either one of her cares or one of her commitments. Not only does it sometimes intuitively make more sense to describe some (attributable) behaviour as having its source in a care as opposed to an evaluative judgment or vice-versa, but the availability of both allows for greater flexibility in our moral reasoning. This is clearest in cases of apparent internal conflict, such as the Huck Finn case or the mother who still loves her serial-killer son . In the latter case, for example, the mother struggles with her evaluative commitment that tells her to reject her son and her care for him that pushes her towards forgiveness. Both attitudes seem praiseworthy, and moreover there is something praiseworthy in the mother's experiencing both of them: her evaluative commitment towards the value of human life causes her to revile her sons actions and makes her a good person; her care for her son causes her to want to forgive him and makes her a good mother. Certainly it would be possible to describe both in the language either of cares or of evaluative commitments, but there is a sense in which that fails to capture the conflict going on in such cases.

To offer a very tentative hypothesis, it would seem that the existence of both characterisations of the deep self allows us to make attributability judgments in a framework of both collective normative demands (morality) and individual ideals. The idea of evaluative commitments most easily captures those underlying attitudes that conform with shared normative expectations, as the 'evaluative' aspect implies conscious recognition (on the part of the agent) of what is (from the perspective of the observer) 'objectively' best. (Where people are exposed to a given moral framework, they are also more likely to *consciously* hold the relevant evaluative judgments as normative propositional beliefs). The idea of cares more easily captures the idea of an individual ideal, as what an agent *cares* about carries with it no necessary implication that anybody else could be expected to do the same. Perhaps that explains not only

the existence of both categories, but also the relative dominance of 'cares-based' approaches in a post-virtue world of individual ideals where 'authenticity' is often seen as a virtue in itself.[66]

Ultimately however the explanatory flexibility provided by the combination of two traditions only completes the argument against them: it supports the hypothesis that philosophical accounts of the deep self do nothing more than formalise our folk-psychological repertoire of rationalisations for our moral responsibility attributions. Moreover, in the process of eliminating the various candidates for a response-independent feature of the deep self capable of supporting the realist assumption, we have begun to see the influence on our attributability judgments both of general cognitive biases (psychological essentialism) and of specific normative judgments. It is with the precise mechanism of these influences that Chapter III, drawing on the empirical literature on the topic, will be concerned.

[66] See (MacIntyre, 1984) and (Taylor, 1992).

Chapter III: The Empirical Results

§1) Introduction

In the first chapter I introduced the concept of the deep self as a feature of our moral responsibility reasoning and presented two alternative ways of approaching it. The Realist approach was to claim that deep selves *exist*, having the response-independent properties that feature in our deep self talk, so as to (objectively) explain and justify our (correct) moral responsibility attributions. This 'natural psychological kind' approach to the deep self suggests not only that it can be characterised in purely descriptive terms, but that what our deep self concept identifies actually plays a deep explanatory role within human psychology – that the distinction we make between those motivations that are 'internal' to an agent's deep self and those from which she is alienated identifies a naturally significant division. The alternative approach – which I described as involving a 'Strawsonian reversal' – was to claim that deep selves are constructed, existing only as constructs within our folk-psychology with no corresponding natural kind; that they were constituted only by response-dependent properties such that there could be no 'fact of the matter' about an agent's deep self.

In the previous chapter I presented a range of 'traditional' compatibilist theories of moral attributability – 'traditional' in the sense that they shared a common commitment to what I have described as the 'realist assumption' concerning the deep self. An analysis of these existing deep self theories, and their development over the last fifty years, revealed an important trade-off. On the one hand, theories that presented a psychologically descriptive (response-independent) criterion for identifying 'deep' (attributable) motivations inevitably failed to capture important instances of either attributable or unattributable desires or evaluative judgments. On the other hand, theories that operated in terms of much vaguer concepts like 'cares' or 'commitments', whilst capable of covering the full range of our possible attributability intuitions, ultimately failed to identify a properly response-independent property of the deep self. Accounts of this second kind, which form the mainstream of contemporary compatibilist approaches to attributability, end up being very good representations of our folk-psychology of the deep self: they provide concepts with little or no descriptive content, which can be employed flexibly so as to rationalise attributability judgments as necessary.

In the absence of any response-independent criterion identifying a natural psychological kind as the object of our deep self intuitions, the response-independent hypothesis starts to

become more plausible. A range of factors were observed, in the previous chapter, as apparent influences on our attributability (and therefore also deep self) judgments or intuitions. Particular features of our folk-psychological or even normative frameworks seemed to determine our deep self attributions in ways inconsistent with their reliably identifying some natural kind. The primary aim of this chapter, in turning to the field of experimental moral philosophy, is to build up a more comprehensive understanding of these various influences in order to fill out what has so far been only the outline of a Strawsonian anti-realist account of the deep self. More specifically, my consideration of a range of empirical studies bearing on the deep self concept has three objectives in mind. The first is to provide empirical support for the range of hypotheses presented in Chapter II concerning variability in our deep self intuitions caused by apparently non-evidential factors. The second is to address an emerging hypothesis relying on deep self realism as an antidote to sceptical concerns arising from the field of experimental philosophy. The third is to present an alternative solution to those sceptical concerns in the form of a unified (anti-realist) explanation of the empirical results bearing both directly and indirectly on the deep self concept.

The unified explanation that I have in mind will involve presenting the deep self concept (and our deep self attributions) as the products of a much more general cognitive mechanism: our causal cognition. I will argue that a feature of this causal reasoning is the postulation of (unobserved, 'deep') causes where it is expedient to do so in order to create 'causal models' of our environment with maximal explanatory and predictive power – that is, in a way that is *computationally rational*. What the attribution of a particular motivation to an agent's deep self would amount to, on this account, is the creation of a causal model in which a given behaviour is explained not by a range of potentially 'causal' features of the agent's situation, but by a postulated 'essence' within that agent that disposes her to act in that particular way. What I will then attempt to show, within this framework, is how the various 'influences' on our attributability judgments operate within this general cognitive mechanism to explain the variability of our deep self intuitions.

In order to get there though, I will first briefly consider the role of empirical approaches in philosophy in general and moral psychology in particular, presenting the general form of a number of sceptical challenges raised by the findings of experimental philosophy and outlining what a potential solution (in the form of a Strawsonian anti-realist approach) might look like [§2]. I then turn to a more specific set of experimental findings that have collectively contributed to what is now known as the 'person-as-moralist' paradigm in moral psychology.

The general observation emerging from these results is that a number of supposedly descriptive concepts, many of which we think of as underpinning moral or normative claims, themselves appear to be subject to a 'normative asymmetry'. I explain why the scepticism arising from such results – often described as the 'bi-directional thesis' – is not so easily resolved by the deep self realism suggested by a number of theorists, and therefore remains in need of an explanation [§3]. Finally I turn to a series of relatively recent studies that address the deep self concept directly. I engage with the mainstream interpretation of these initial results – a hypothesis best described as 'deep self optimism' – before suggesting how, by applying the kinds of explanatory tools derived similar empirical analyses of philosophical concepts, a more satisfying, unified explanation can be provided in the form of my Strawsonian anti-realist approach.

§2) X-Phi: Sceptical Challenges and Potential Solutions

My aim in this section is to interrogate briefly the relationship between traditional philosophical methods and the empirical methods that are the tools of the experimental philosopher. Obviously the emergence of experimental philosophy in general, and moral psychology in particular, has led to some not insignificant controversy concerning the place of empirical methods (their authority, their interpretation, their reliability) within traditional philosophical theory.[67] My relatively modest ambition, in this chapter, is to provide an outline of the kinds of challenges that experimental philosophy (X-Phi) might pose to the kinds of theories that I have been discussing so far: theories founded on some form of the realist assumption, and committed, moreover, to a kind of reductive naturalism. This is one particular way of doing moral philosophy that assumes that the truth of normative claims is dependent on features of the world that are reducible to particular descriptive facts about the subjects of those normative claims. So, for example, the kind of reductive naturalism that we have seen so far with respect to deep selves is the suggestion that a claim within our normative discourse – that property *x* is attributable to agent *A* for the purposes of moral responsibility – has its truth conditions in some descriptive facts about that agent's psychology.

[67] To the extent that critiques of experimental philosophy address specific methodological issues – raising concerns, for example, with the widely used survey methodology – these should be borne in mind as a general caveat to the interpretation of empirical results, just as it is in fields like social psychology in which the survey methodology is well-established.

Remaining, however, with the general form of reductive naturalism and the realist assumption that goes with it, I want to explore three general forms of sceptical challenge that experimental philosophy might present in order to then compare two different ways in which one might respond to these sceptical challenges – one that involves retaining the realist assumption, and the other that involves adopting a Strawsonian anti-realism. The sceptical challenges facing the traditional realist approach are the following:

(i) The Sceptical Challenge to the Universality of Intuitions

One significant project within experimental philosophy, most prominently advanced by Weinberg and Stich,[68] aims to undermine what is seen as an ultimately empirical claim being made in traditional philosophical conceptual analysis: that the philosophical intuitions being used as 'data' for such an analysis are widely shared enough to be considered universal. Such traditional analyses reveal a commitment to what Weinberg and Stich (pejoratively) describe as 'Intuition-Driven Romanticism': the idea that the truth of a particular concept is contained within the philosopher's mind and needs only be expressed by means of appropriately elicited intuitions.[69] For reductive naturalists this means examining intuitions about something like 'attributability' (that is, examining attributability judgments in a range of situations) in order to discover the descriptive properties to which those normative intuitions are responding. The concepts in question need not be strictly 'moral' either: epistemic intuitions about what qualifies as 'knowledge' appear to vary significantly along cultural and socioeconomic lines (Weinberg, Nichols, & Stich, 2001), semantic intuitions suggest that the causal-historical view of reference is more likely to be found among Westerners than East-Asians (Machery, 2004) and intuitions to a range of well-known philosophical dilemmas appear to differ along gender lines (Buckwalter & Stich, 2014).[70]

So for example, as raised in the previous chapter, the possibility of individuals with radically different frameworks having opposing attributability intuitions would raise serious questions as to the universalizability of any particular attributability intuition that might feature as a 'datum' in a theory of the deep self. Even if philosophers in general *agree* in their

[68] See e.g. (S. Stich & Weinberg, 2001; Weinberg, Nichols, & Stich, 2001).

[69] See (Weinberg et al., 2001)

[70] Whilst these results apply specifically to 'intuition-based' responses, there has been a much more general realization within experimental psychology concerning the problem with drawing conclusions about 'human psychology' from WEIRD people (subjects drawn from Western, Educated, Industrialised, Rich, Democratic societies, typically American undergraduate students) (Henrich, Heine, & Norenzayan, 2010).

attributability intuitions, this only raises a further question concerning philosophical 'expertise'.[71]

(ii) The Sceptical Challenge to the Reliability of Intuitions

A further significant project within experimental philosophy looks to challenge the reliability of philosophical intuitions independently of whether such intuitions can be shown to be universal or not. A useful summary of the general proposition comes in the form of an answer to the question 'when does psychology undermine belief?' and is as follows:

> A belief p may be undermined whenever: (i) p is evidentially based on an intuition which (ii) can be explained by a psychological mechanism that is (iii) unreliable for the task of believing p; and (iv)any other evidence for the belief p is based on rationalisation (Leben, 2014, p. 328)

In other words, if 'non-evidential' factors can be shown to have an influence on people's intuitions concerning a particular concept, we would have reason to treat those intuitions (or rather, the psychological mechanisms that produce them) as unreliable. A psychological mechanism might unreliable for the task of believing that *p* (where *p* is, for example, a proposition concerning attributability) if it were subject to some kind of 'cognitive bias'. One famous example is the influence of 'framing effects' on responses to moral dilemmas first explored by Tversky and Kahneman (1981).[72] The general form of the sceptical concern arising from results such as these is the following: an intuition that *p* (where *p* might be an intuition that one of two normative claims is true) is often taken as evidence for *p*, yet within a framework

[71] The questions to be answered are (i) whether philosophers' intuitions are significantly different to those of 'the folk', and (ii) whether, if they are, there is any reason to think that philosophers' intuitions are to be considered *more* reliable than those of the non-philosophical folk (see: Andow, 2015; Liao, 2016; Mizrahi, 2015; Ryberg, 2013; Schulz, Cokely, & Feltz, 2011; Schwitzgebel & Cushman, 2012, 2015; Tobia, Buckwalter, & Stich, 2013; Weinberg, Gonnerman, Buckner, & Alexander, 2010; Williamson, 2011). Empirical results concerning differences in intuitions between philosophers and non-philosophers are largely preliminary and variable depending on the particular intuitions under investigation, with some studies concluding that philosophers' intuition are subject to the same influences as those of non-philosophers (see e.g. (Schulz et al., 2011) on judgments concerning free-will and moral responsibility and the influence of extroversion), and others appearing to show differences in intuitions between philosophers and non-philosophers, but questioning whether or not there is reason to believe that such differences may be interpreted as showing the philosophers' intuitions to be *more* reliable (see e.g. (Tobia et al., 2013) on the actor-observer bias in judgments of moral permissibility and obligation). Ultimately, the question of the *reliability* of intuitions generally forms the subject matter of a second sceptical concern, to be examined below.

[72] The 'framing effects' in question involve simple changes in the presentation of choice problems – e.g. the presentation of alternative programs to combat a disease in terms either of the number of people each is predicted to save as opposed to the number of people each is predicted to kill – which can lead to significant changes in people's response tendencies (in the case in question, in the 'save' condition participants were more likely to take a risk to potentially save more people, whereas in the 'die' condition participants were more likely to avoid risk, when the average expected outcome of each program in reality remained the same). Later contributions have applied similar methods to more traditional philosophical problems such as trolley cases (Petrinovich & O'Neill, 1996) or passing moral judgment for an action as opposed to an omission (Haidt & Baron, 1996).

of traditional reductive naturalism (where the truth of *p* depends on certain descriptive properties of the world) the truth of *p* cannot depend on something as subjective as the order in which the two options are presented. And yet, empirical results suggest that people's intuitions often *do* depend on the order in which options are presented, or their wording, or any number of other factors that are clearly 'non-evidential' in the sense that they have no bearing on the descriptive properties of the subject matter of the intuition.[73]

(iii) The Sceptical Challenge to the Role of Moral Reasoning

A final kind of project emerging from experimental philosophy that is worth mentioning here concerns the link not between intuitions and normative 'truths', but between moral reasoning and *psychological* truths. Specifically, it aims to challenge the validity of introspection as a reliable way of understanding how our minds work – and what exactly they are doing (what descriptive features of the world they are responding to) when they produce things like intuitions. This kind of sceptical concern is of even greater relevance to descriptive naturalists who would claim that the descriptive properties underpinning many normative truths are in fact psychological properties. So, for example, the method of the deep self realists discussed in chapter II involves two potentially problematic kinds of introspection: the introspection of the folk-psychologist, who claims to perceive that the reason for a particular attributability intuition is that the behaviour in question is expressive of an agent's deep self, and the introspection of agents themselves insofar as they present explanations for their behaviour in terms of 'deep' motivations like cares or commitments.

The starting point for the sceptical worry about moral reasoning or introspection is the observation of a kind of 'incongruence' between the reasons that agents might present for their

[73] Whilst this will be developed later with a range of 'non-evidential factors' that influence attributability judgments, it is certainly true that the argument with respect to simple 'framing effects' does not necessarily lead to a sceptical response. One possible response to this kind of debunking argument with respect to framing effects (see for example the more recent debunking projects of (Sinnott-Armstrong, 2008) and (Nadelhoffer & Feltz, 2008)) is simply to dispute that the observed effects are significant enough to warrant scepticism. This is the approach taken by (Demaree-Cotton, 2016), who argues, based on a meta-analysis of framing effect-style empirical results, that the figures simply do not warrant the conclusion that our moral intuitions are unreliable. What he claims is that the empirical results on average show our moral intuitions to be 80% reliable, and subject to some framing effect which is however not determinate. The extent of the framing effect – or the rate of unreliability – is not sufficient, according to the author, to defeat the argument for non-inferential justification. This makes sense when we consider that many of the more significant effect sizes (such as those obtained by (Tversky & Kahneman, 1981)) involve participants being forced to choose between two essentially equivalent answers (recall that in the 'disease-program' case, both options had the same average expected outcome), so while the effect itself may be noticeable, it is not clear that it is sufficient to reliably 'make up an agent's mind' with respect to options that are not similarly equivalent.

actions and those that we might describe as the *actual* reasons (the causes, perhaps) of those same actions. In short, it seems possible that the actual mechanism by which we arrive at (for example) normative intuitions or judgments is opaque to us, and our conscious reasons-based explanations might be nothing but a post-hoc rationalisation. The evidence for this hypothesis comes from a range of sources, most notably summarised by Jonathan Haidt (2001). His arguments in support of scepticism about moral reasoning include the following: (i) psychological evidence of 'dual process systems' suggesting that we have both a default (automatic) evaluative system for reaching moral judgments quickly (relying on a number of simple heuristics) and a more complex cognitive ability to consciously reflect on reasons; (ii) evidence concerning the role of 'motivated reasoning' in human cognition, suggesting that the slower, conscious reasoning process is primarily activated not as an alternative to the rapid intuitive/perceptual process, but in order to provide explanations for the conclusions already reached by that system (with an emphasis on providing explanations that both facilitate social interactions – the 'relatedness motive' – and avoid cognitive dissonance – the 'coherence motive'); (iii) a number of experimental studies in which subjects are seen to present explanations for their behaviour in terms of reasons that could not have been available to them at the time.

So for example, with respect to any given attributability intuition, one concern might be that our introspection as to the reasons behind the intuition is not detecting the actual *causes* of the intuition at all but instead providing a mere rationalisation of it. The same might go for agents' introspection concerning their own behaviour:

> When asked to explain their behaviours, people engage in an effortful search that may feel like a kind of introspection. However, what people are searching for is not a memory of the actual cognitive processes that caused their behaviours, because these processes are not accessible to consciousness. Rather, people are searching for plausible theories about why they might have done what they did. People turn first to a "pool of culturally supplied explanations for behaviour," which Nisbett & Wilson (1977) refer to as "a priori causal theories". When asked why he enjoyed a party, a person turns first to his cultural knowledge about why people enjoy parties, chooses a reason, and then searches for evidence that the reason was applicable. […] If this reverse path is common, then the enormous literature on moral reasoning can be reinterpreted as a kind of ethnography of the *a priori* moral theories held by various communities and age groups. (Haidt, 2001)

Certainly one can already begin to see the potential problems this might pose for deep self realists. If the first two sceptical concerns turn out to apply to people's deep self intuitions and attributability judgments – that is, if these intuitions are not universal, and are influenced by factors that might be described as 'non-evidential' in the sense that they do not relate to

descriptive properties of the object of the intuition – then it seems unavoidable that the explanations we commonly present for such intuitions (both as philosophers and as folk-psychologists) will not reflect their actual *causes*.

Returning to the general form of these sceptical challenges, I would like to sketch two alternative ways in which one might potentially respond. The first option – a 'performance error' account – is the realist option, because it involves retaining a commitment to the idea that the truth of normative claims that are the subject of such sceptical challenges is nonetheless dependent on features of the world reducible to descriptive facts about the objects of the normative claims. What this option inevitably requires is some kind of explanation as to how so many of our intuitions or normative claims turn out to be wrong – how, for example, if a certain moral intuition is not shared between different cultures, at least one of those groups must be failing to accurately identify the underlying moral facts, or how our intuitions, as mechanisms by which we identify certain moral facts, could so often be influenced by 'non-evidential' factors. What this typically involves is an account of some kind of 'interference': a reason why, even though what our normative intuitions are trying to do is track certain objective (descriptive) features of the world, there are other factors that persistently interfere with that mechanism. In this way any variability in people's intuitions would ultimately be explained in a way that involved at least some of those people getting it *wrong*, and any even universal susceptibility to 'non-evidential' factors would require an explanation as to how everyone, reliably gets things wrong in certain situations. The resulting theory would not be a 'moral error theory' – because it would not assert that all normative claims are false – but what has been described as a 'performance error' explanation: one that retains its commitment to the possibility of true normative claims, but explains why our intuitions sometimes (or even quite often) fail to point us to them.

The potential problems with such an approach are numerous. Take, for example, the initial empirical investigations into folk intuitions concerning compatibilism/incompatibilism (discussed in Chapter I). Empirical results seemed to show that when presented with an abstract description of a determinist universe, subjects would reliably produce incompatibilist moral intuitions (i.e. would not blame agents in such a universe) but when presented with concrete scenarios of moral wrongdoing would produce compatibilist intuitions (i.e. would blame agents

in the fictional scenarios).[74] As pointed out by the researchers, there are a range of potential interpretations one could have of these results. The two clearest alternatives are the following: (i) people understand that moral responsibility is incompatible with a determinist universe, and are able to express this understanding when presented with abstract scenarios, but when presented with concrete scenarios involving moral wrongdoing experience negative affect, which in turn elicits punitive tendencies and so ultimately a judgment (as the expression of these punitive tendencies) that the agent in question (despite the determined universe in which she is operating) *is* morally responsible; (ii) people's concept of moral responsibility is insensitive to the truth of determinism (i.e. compatibilist) as represented by their immediate intuitions to concrete scenarios, but when presented with an abstract scenario they suffer from 'theory contamination' such that their immediate, perceptual (intuitive) response is overridden by an abstract propositional belief that freedom is required for moral responsibility. Both alternatives can be (and have been) argued, but the point is that neither is really *supported* by the evidence so much as it finds a way to make the evidence *fit* a pre-conceived compatibilist or incompatibilist conclusion. It is a profoundly unsatisfying way of dealing with empirical results.

The second option is what might be described as a 'conceptual competence' account of the data, which involves explaining the results in a way that doesn't assume that some not insignificant proportion of people get things wrong some not insignificant amount of the time. In order to achieve this – to find ways, for example, in which the influence of supposedly 'non-evidential' factors on agents' normative intuitions represents not a mistake but a basic competence with the concepts in question – our perspective will need to undergo a Strawsonian reversal. What this means is that instead of our intuitions being relatively unreliable indicators of some elusive, underlying category of objective moral facts, they should be understood from an anti-realist perspective, either as quite reliable indicators of moral facts relative to specific observers, or otherwise simply as expressive of certain facts about the different folk-psychologies from which they emanate. What it will be interesting to explore is precisely which elements of agents' natures affect what is true for people like them. So for instance instead of attempting to explain the claim that *x* is attributable to an agent based on descriptive features of that agent's psychology alone, we can attempt to explain why that proposition is true from the embedded perspective of the person making the judgment. Just how such explanations function

[74] See (Nichols & Knobe, 2007).

in their application to results from experimental moral psychology remains somewhat vague in the abstract, and so the next section will aim to provide an extensive set of examples of what 'conceptual competence' accounts look like.

Before moving on to the next section though, I would like to make one final note concerning what is problematic about 'performance error' accounts and the kind of genealogical debunking of normative intuitions that they involve. I say 'genealogical debunking' because from the realist perspective, according to which normative claims are objectively true or false based on descriptive properties of the world, any intuition or judgment that has a genealogical explanation in terms of some feature of collective or individual psychology (as opposed to responding more or less directly to the descriptive properties on which its truth depends) must be found to be 'non-evidential' with respect to the normative claim in question. Take, as an example, Greene's (2008) 'empirical debunking' of deontological morality. His claim is that deontological morality is ultimately rationalising intuitions that, rather than tracking moral 'facts' about the world, can be explained with reference to "amoral evolutionary causes", or "morally irrelevant factors having to do with our shared evolutionary history" or "morally irrelevant constraints placed on natural selection in designing creatures that behave in fitness-enhancing ways" (Greene, 2008). In short, according to Greene, "the natural history of our retributivist dispositions makes it unlikely that they reflect any sort of deep moral truth". If a moral intuition can be explained in genealogical terms then it is unreliable because it is the product of 'morally irrelevant factors', but what is left rather undeveloped is what exactly a 'morally relevant' factor might look like – one, no doubt, that has no roots in human nature or evolutionary history whatsoever but in some way responds directly to 'normative facts'. Greene is aware, of course, of the problem with this line of argumentation – recognising that to follow such a sceptical project to its conclusion would be to obliterate human morality altogether – but tells us that the problem (of where to draw a line in our 'genealogical debunking') is one to be dealt with elsewhere.

What Greene suggests is that consequentialist principles, as the best standards for public decision-making, can tell us "which aspects of human nature it is reasonable to try to change, and which ones we would be wise to leave alone". In doing so of course he inevitably introduces a new kind of normative intuition that is equally likely to fall prey to the same genealogical debunking arguments: whatever we judge to be of 'value' or 'consequence' or 'utility' as the foundation of a given set of consequentialist principles is likely to be as much a product of our human natures and 'amoral evolutionary history' as the 'retributivist' intuitions that he railed

against. Ultimately such a sceptical project leaves the Realist with nothing to work with, having 'kicked away the ladder' of moral intuitions in search of objective normative truths. This is why, in what follows, I will formulate an alternative approach that avoids Greene's problem. It is an approach that, rather than rejecting a whole range of moral intuitions as 'performance error' – for failing to identify some independent set of moral facts – takes those intuitions as the necessary starting point for determining the 'moral facts', perhaps even as constitutive of them.

§3) Conceptual Competence in the 'Person-as-Moralist' Paradigm

Let us begin by considering a now well-known example of experimental moral psychology known as the 'Chairman Experiment'. In it, subjects were presented with the following alternative vignettes:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but (/and) it will also harm (/help) the environment.'
>
> The Chairman of the board answered, 'I don't care at all about harming (/helping) the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed (/helped).
>
> (Knobe, 2003)

These two alternative vignettes elicited strikingly different patterns of response – an asymmetry now commonly referred to as the 'Knobe Effect'. In the 'harm' condition, 82% of subjects responded that the side effect (harming the environment) was brought about intentionally; in the 'help' condition, 77% of subjects responded that the agent *did not* bring about the side effect (helping the environment) intentionally. And yet, on the surface at least, the Chairman's attitude towards each side effect is exactly the same: he couldn't care less what the side-effect is, and doesn't turn his mind to it in the least. Are test subjects making some error when they attribute an intention in the one scenario but not in the other, or could the concept of intentional action somehow incorporate this apparent 'normative asymmetry', or valence effect? The important (and potentially worrying) possibility that results like these raise has been described as the 'bi-directional thesis': our normative judgments (for example, that it is good/bad to help/harm the environment) as moral evaluations – a kind of claim about what *ought* to be the case – appear to be influencing our factual or descriptive judgments about what *is* the case (about an agent's intentions). This would be especially problematic in cases where

the factual or descriptive claims in question supposedly justify normative claims about something like blameworthiness. This kind of bi-directional relationship between normative and descriptive claims risks undermining the justification of normative intuitions/judgments on the basis of the supposedly descriptive facts on which they depend.

Widespread observations of the ways in which normative intuitions and judgments seem to play a role in our use of apparently descriptive concepts have been described as the 'person-as-moralist' paradigm in our understanding of folk-psychology and causal cognition (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Knobe, 2010). To take a few simple examples, we would ordinarily assume that what it is to be *happy*, or to *value* something, or to be *in love* with someone involves a particular kind of mental state, such that the truth of any particular claim ('that *A* is happy'; 'that *A* values *x*'; 'that *A* is in love with *B*') would be determined by the (objective) existence of the relevant mental states in the agent in question (happiness, valuing *x*, and loving *B* respectively). And yet experimental investigations into these concepts reveal a normative asymmetry in attributions of happiness, valuing or love that does not, on the face of it, appear to be simply tracking the mental states of the agents in question.

Knobe and Preston-Roedder (2009) investigated the folk concept of 'valuing' by presenting subjects with vignettes like the following:

> Susan grew up in a religious family, but while she was in college, she started questioning her religious beliefs and eventually became an atheist.
>
> She will be getting married in a few months to her longtime boyfriend. Recently, the subject of premarital sex has come up.
>
> Susan definitely has a desire to have sex with her boyfriend, but whenever she thinks about doing so, she remembers what he church used to say about premarital sex and feels terribly guilty. As a result of these feelings, Susan has not had sex yet.
>
> Because she is no longer religious, Susan believes there is nothing wrong with premarital sex. She wishes she could stop feeling guilty and just follow her desires.
>
> (Knobe & Preston-Roedder, 2009, p. 135)

Subjects from two different groups (members of a Mormon bible-study group and passers-by in a public park) were asked (i) whether Susan 'valued' refraining from premarital sex, and (ii) whether doing so was good, bad or neutral. The Mormons, judging such actions to be good, overwhelmingly answered that Susan *did* value refraining from premarital sex,

whereas the passers-by (judging such actions to be neutral) generally answered that she did not.[75]

The authors also report unpublished data from Nyholm (2007) on the concept of happiness, involving doctors in either a Red Cross hospital or a Nazi death camp who both get some pleasure from their work, but equally find it difficult and sometimes upsetting, though they both ultimately think it is an important thing to do and feel a corresponding sense of satisfaction. Despite the symmetries in their mental states, participants reliably agreed that the Red Cross doctor was happy, but denied that this was the case with the Nazi doctor. Phillips, Misenheimer and Knobe (2011) report a similar asymmetry with respect to the concept of 'love'.[76]

What is significant about the asymmetries observed is that they are not the products of simple 'framing effects' (that might more easily be explained away as a kind of 'performance error') but instead involve the framing of the situation – and indeed all of the psychologically descriptive elements – being held constant whilst it is the moral valence of the situation that changes. This presents obvious problems for what Knobe (2010) describes as a 'person-as-scientist' understanding of folk psychology: clearly if folk-psychologists were primarily interested in accurately describing the world around them, they would not employ such problematic 'descriptive' concepts as these.[77] Realists about the psychological properties in question – those who would insist that there is some 'natural kind' (mental state) that constitutes, for example, 'happiness' – would be forced to accept that the concept of 'happiness' that they have in mind is something distinct from the concept that features in our folk-psychology.[78]

[75] Interestingly, these experiments also contrasted the concept of 'valuing' with the concept of 'thinking good', with the result that the concept of 'thinking good' was *not* subject to the same normative asymmetry.

[76] In each case the 'normatively mediated' concept ('happiness' or 'love') was compared with an alternative concept ('unhappiness' or 'lust') that showed none of the normative asymmetry. Apparently, whilst subjects are unwilling to say that a subject leading what they deem to be a 'bad' life is happy, they had no problem agreeing that an agent leading what they deem to be a 'good' life is unhappy.

[77] Alicke and colleagues (2015) present a more detailed account of different 'metaphors' of causal cognition, from 'person-as-scientist' (in which agents attempt to describe the world around them in objective terms) to 'person-as-lawyer' (with an emphasis on motivated cognition and the use of causal reasoning to inculpate/exculpate certain individuals) to 'person-as-moralist', where people's causal attributions appear to be primarily expressive of moral condemnation.

[78] Certainly some philosophers may be happy to say that the concept of happiness is normatively loaded rather than merely descriptive of some natural kind of mental state. But it is not enough to say, for example, that our 'happiness' attributions, in addition to their descriptive role with respect to a particular kind of mental state, are *also* expressive of a normative judgment – to do so is simply to accept that our ordinary concept of happiness, insofar as it describes a natural kind, is subject to some kind of interference. The kind of conceptual competence account I am proposing would involve denying that there *is* any natural kind of mental state identified (even

Where the normative asymmetries in question become even more problematic though – and where we start to see the emergence of the bi-directional thesis – is with respect to certain seemingly descriptive concepts that are supposed to underpin important normative claims. Our attributions of moral responsibility (before even arriving at the question of deep selves) in many cases rely on combinations of the following 'descriptive' claims: that the agent in question was *causally* responsible for the outcome in question; that she *intended* to do the action in question (again, in the most straightforward sense of intentional action, not yet involving questions of 'deep' motivations); and that she *believed* or *knew* that her actions would have the relevant outcome. So, for example, in order to blame the Chairman for harming the environment, we would need to argue that some actions of his were causally responsible for the harm in question, that they were intentional, and that he knew or believed that those actions would lead to the harm in question. In order to examine what is so potentially problematic about normative asymmetries in our folk-psychology and causal cognition, I therefore propose to consider those three folk concepts of (i) causation, (ii) intentionality and (iii) knowledge. In each case I will examine what a 'performance error' account might look like – one that operates by reference to an independent fact of the matter in each case and therefore involves at least some subjects reliably being *wrong* about it – and compare it with an alternative 'conceptual competence' account that suggests that subjects are, from an embedded perspective, getting things *right* in their application of the relevant concepts.

(i) Causation

Consider the following experiment. Subjects are presented with a vignette in which an agent, driving above the speed limit, is involved in a crash. They are then asked to identify the *cause* of the crash, where the role of the agent's speeding is compared to a range of alternative causes in different versions of the vignette (an oil spill that makes it impossible for him to stop in time; a tree branch obscuring a stop sign at an intersection; another car running through a stop sign). The primary variable across different vignettes was the 'moral valence' of the driver's actions: in one condition he is speeding home to hide an anniversary present for his parents before they arrive; in another he is speeding home to hide a vial of cocaine that he has left out in his room. In the experiment, subjects were significantly more likely to identify the

unreliably) by our happiness intuitions – a bullet that becomes even harder to bite when it comes to concepts like intentional action.

speeding driver as the cause of the crash when his motive for speeding was blameworthy – that is, in the 'cocaine' condition (Alicke, 1992).[79]

From a scientific perspective we would be inclined to say that the blameworthiness of the agent's motive is clearly irrelevant to the causal contribution of his speeding to the crash that ultimately results, and then search for explanations as to why our causal intuitions seem to get something wrong. The 'performance error' account of results like these – that fits into the 'person-as-moralist' framework – might suggest that the negative affect experienced towards the 'cocaine-hiding' driver leads to a judgment of blameworthiness which is then rationalised by the causal claim that the driver in question was a more significant *cause* of the crash.

An alternative explanation is presented by Hitchcock and Knobe (2009) according to which the relationship between norm violations and causal attributions is perfectly rational when our causal cognition is considered in its functional context. What is meant by this 'functional context' is the fact that our causal cognition is not solely (or even primarily) devoted to representing descriptive facts about the world. Its purpose, instead, is to identify possible points of intervention that allow us to *manipulate* the world around us.[80] In this context, it is highly rational that we should attribute causal influence to elements of the world that are 'norm-violating': these are likely to provide the best general rules for future interventions to avoid the negative outcome in question. Consider a car crash at an intersection where one car is speeding and the other respecting the speed limit: it makes sense in cases like this to say that the speeding driver *caused* the crash even though both drivers needed to be driving at their precise actual speed in order for the crash to come about (both are necessary conditions for the crash in question). Even though it is true that if the *other* car had been speeding then the crash would never have occurred (because the two cars would not have crossed the intersection at the same time) this is not particularly helpful. Even though it would have worked in this very particular case, with the other driver approaching the particular intersection at exactly the speed that he was travelling, the general rule 'drivers should speed' is not a very reliable way of avoiding crashes in future.[81] Sensitivity to norms like these "*directs our attention away from*

[79] Knobe and Fraser (2008) show similar results where the 'moral valence' variable is replaced with a 'norm-following' v 'norm-violating' variable, where the norm in question involves a non-normatively charged rule about administrative assistants being allowed to take pens and professors not.

[80] There are numerous philosophers who have defended the claim that causal properties simply *aren't* purely descriptive properties on the world, but something more like 'secondary qualities' that can only be described by reference to human agency and causal cognition. See, for example, (Menzies & Price, 1993).

[81] The same thing is happening, in a more indirect way, in the car-crash example presented by Alicke (1992). Subjects must choose between various norm-violating elements of the situation, comparing a driver's speeding

interventions that work by leaving in place some highly unusual aspect of the situation and then capitalising on it to achieve a particular effect" (Hitchcock & Knobe, 2009). Sensitivity to moral norms in particular – as revealed in this experiment – is a way of optimising interventions such that they involve improving on some 'bad' aspect of a situation rather than altering or removing some 'good' aspect.

Within this functional context of our causal cognition, it can hopefully be seen how the tendency to identify norm-violating features of situations as causally responsible for unusual consequences does not involve any kind of 'error' in our causal intuitions at all. Instead it represents the perfectly correct application of a concept that is itself not 'scientifically descriptive' (in the sense of trying to identify some independent 'natural kind' that is causation) but rather functional for creatures like us, as a way of producing generalisable rules for the future.

(ii) Intentional Action

The basic outline of the intentional action asymmetry – also known as the 'side-effect effect' – has already been given above in the form of the Chairman experiment.[82] Two primary 'performance error' explanations have been put forward for the general phenomenon of asymmetrical intentional action ascriptions: 'conversational pragmatics' and 'theory interference'. The former suggests that participants don't *actually* judge the effects in question (expected but unintended side-effects or intended but only fortuitously achieved side-effects) to have been brought about 'intentionally', but want to avoid the unwanted conversational implicature that might follow from such a statement – that is, that the agent in question is not

with, for example, an obscured stop sign or another driver's negligence, where both would create reasonable general rules for avoiding crashes ('don't speed' or 'don't have obscured stop signs'). The introduction of the agents' intentions – as causes *behind* their speeding – changes the calculation somewhat: the general rule 'don't speed in order to do good things' (like doing something nice for one's parents) will lead to less positive outcomes than the rule 'don't speed in order to do bad things' (like hiding cocaine) and so the difference in the drivers' intentions accounts for the norm-following agent's behaviour being slightly less likely and the norm-violating agent's behaviour being slightly more likely to be prioritised over the competing causes.

[82] It is worth noting that the findings of the initial Chairman experiment have since been significantly developed and reinforced/confirmed. For example, drawing on the observation that attributions of intentionality generally require that the agent in question have the necessary *skill* to reliably bring about the outcome in question (Malle & Knobe, 1997) Knobe, in a separate paper (2003b), observes that this feature of intentionality judgments is *also* subject to the normative asymmetry: when it comes to 'good' (praiseworthy) actions, whether or not the agent in question had the skill to reliably bring about an outcome is an important determinant of the judgment that she *intentionally* brought about that outcome; when it comes to 'bad' (blameworthy) actions, the skill of the agent performing the action makes much less of a difference, and participants are likely to judge the outcome as brought about intentionally regardless. The basis intentionality asymmetry has also been replicated across cultures (Knobe & Burra, 2006) and at different stages of development (Leslie, Knobe, & Cohen, 2006).

to blame for them. The idea then is that the claim that such consequences are brought about intentionally is really a way of trying to express condemnation towards the agents for their behaviour and assert that they are to blame for its consequences. The latter explanation suggests that while participants' 'intuitive' responses to the scenarios are that the consequences are not brought about intentionally, they hold a propositional, theoretical belief that in order to be blame an agent for something it must have been brought about intentionally. The participants would then retrospectively correct their intuitions (modifying their intentionality judgment) in order to maintain their theoretical beliefs concerning the link between intentional action and blameworthiness.

Further elaborations on the intentional action asymmetry suggest that rather than tracking the 'moral valence' of the relevant outcomes specifically, participants' intuitions seem to be sensitive to 'norm violation' in general (meaning that non-morally charged norms might elicit the same asymmetry, or that the asymmetry might be reversed by getting participants to imagine a context of reversed moral norms).[83] The utility of the intentional action concept that is sensitive to norm violations is suggested by Uttich and Lombrozo (2010) who demonstrate in their own experiments that norm-violating behaviour makes a significantly greater contribution to future behavioural predictions than does norm-following behaviour. To understand why this is the case, we have to consider the work that intentional action ascriptions are doing within our folk-psychology and causal cognition: making an intentional action ascription – even for an uncharacteristic action – means attributing to a specific agent a (psychological) disposition to act in the particular way observed. It might not directly translate to a simple behavioural prediction, as the disposition in question may reliably be overcome by situational factors, but it nonetheless becomes part of a model of how we expect another agent to behave, with the expectation that there will be contexts in which it manifests itself.

In the helping chairman's case, the fact that he goes ahead with a lucrative project that is also likely to help the environment (norm-following) provides little to no useful information about how he is disposed to act towards the environment generally: his decision could be

[83] Knobe and Mendlow (2004) first demonstrate that it is not the perceived blame- or praiseworthiness of the agent that explains the asymmetry (as suggested by some of the performance-error accounts) but the normative status of the outcome in question. Knobe (2007) compares judgments of intentionality with participants' readiness to present 'reasons-explanations' for an agent's behaviour, finding that intentionality judgments (which typically feature in judgments of praise and blame, and so operate primarily within the subject's *own* normative framework) might be withheld for agents operating in an alternative normative framework, 'reasons-explanations' (which involve inhabiting another agent's normative framework so as to understand *why* she acted the way she did) are not.

consistent with someone who cares about the environment or someone who doesn't care about it at all. In the harming chairman's case on the other hand, the fact that he goes ahead with a project that is likely to harm the environment *is* informative: it tells us that he clearly cares *less* about the environment than he does about profit, and that he is disposed to disregard the environmental effects of his actions generally – which may serve as a useful piece of information for predicting his behaviour in future.[84] This would be true even if the intentional action in question were an uncharacteristic one.[85] In short, if the intentional action concept is understood as a scientific concept that ought to describe some natural kind of mental states in the agents to which it is applied, then it is easy to see how the intentional action asymmetry would lead to the conclusion that people are getting something *wrong* in their application of the concept. But if the concept is understood as a functional element of our folk psychology and causal cognition – serving to attribute underlying behavioural dispositions – then we can just as easily see how the intentional action asymmetry involves perfectly rational and indeed *correct* attributions[86] in light on the evidential significance of norms and norm-violation.

It is worth pausing, at this stage, to consider an alternative explanation of the intentionality asymmetry alluded to earlier – one that depends entirely on a kind of deep self realism. The basic argument is that the intentionality asymmetry can be explained not in terms of normative differences between alternative outcomes (side-effects, in the Chairman cases) but in terms of participants attributing "*underlying values and characterological dispositions to the agent*" and labelling as 'intentional' behaviour that coincided with those values and dispositions (Sripada & Konrath, 2011).[87] Essentially, the suggestion is that people's intentionality

[84] Uttich and Lombrozo's test participants, for example, were asked to predict how likely a Chairman was to act in a way that either helped or harmed the environment (on a seven point scale) in the future. Some participants were given no information about the Chairman so as to create a 'baseline' of behavioural predictions, and other participants had previously been presented with either the helping or harming chairman scenario. Those presented with the 'helping' scenario predicted harming tendencies slightly above the baseline but still essentially neutral (likely due to the fact that the Chairman's failure to endorse a positive side-effect in the helping scenario is itself a minor norm violation) whereas those previously presented with the 'harming' scenario predicted that the chairman was significantly more likely to act in a way that harmed the environment.

[85] The point is that the judgment in question doesn't involve a simple assessment of how likely an agent is to act in a particular way, but a more specific attribution of an underlying disposition which may or may not reveal itself particularly regularly. Perhaps the chairman ordinarily acts in accordance with various moral and other norms but then uncharacteristically (but intentionally) does something to harm the environment – perhaps because on this occasion he thinks he can get away with it. Whilst we understand that under normal circumstances situational factors will lead him to act in a certain way, his underlying disposition (should those situational factors change) is to act in a different way, and this helps us to make generalisable predictions about how he will act in different contexts where the relevant situational factors might have changed.

[86] 'Correct' in this context refers to a proper application of the intentional action concept, meaning 'functionally correct' and not referring to any independent truth conditions concerning intentional action.

[87] The argument is first formulated in (Sripada, 2010) and repeated in (Sripada, 2012).

intuitions *are* tracking a natural kind, but that that natural kind is actually a kind of 'deep' intentionality, or simply the deep self. On its face the explanation is extremely successful, as our intentionality judgments in cases involving the 'intentionality asymmetry' *do* coincide with our deep self attributions to the agents in those cases. Indeed, the 'deep self' explanation can be extended to a range of similar asymmetries, including the 'valuing' and 'happiness' asymmetries mentioned earlier (Newman, De Freitas, & Knobe, 2015). The problem with relying on this explanation as a way to retain the 'scientific' understanding of the concepts in question (that is, as concepts that identify natural kinds in the world) is that it is far from clear that our deep self attributions are themselves free from the kind of scepticism-inducing irregularities that they are supposed to explain away. Indeed, as we consider, in the next section, the range of apparently 'non-evidential' influences to which people's deep self intuitions are subject and the significant variations that they display, we will be forced to conclude either that the 'deep self' that they are tracking is not a natural kind (as argued in Chapter II) or that some significant amount of deep self intuitions are simply wrong. Whilst such a performance error account may be possible for committed deep self realists, a deep self that people are reliably wrong about would not be capable of displacing the kind of sceptical concerns that the appeal to deep self intuitions is intended to resolve. If, as it turns out, the deep self concept is subject to the same asymmetries as, for example, the intentionality concept addressed here, it will seem much more plausible that the appeal to the deep self is simply a folk-psychological rationalisation of intentionality attributions that are unjustifiable in terms of any 'natural kind' theory of intentionality.

(iii) Knowledge

The knowledge asymmetry, or what has been described as the 'epistemic side-effect effect', stems from a variation on the original Chairman experiment in which subjects were asked whether the chairman *knew* that the course of action embarked upon would help (or harm) the environment (Beebe & Buckwalter, 2010). The results showed that participants were significantly more likely to agree that the chairman *knew* about the side-effect in question in the 'harm' condition than in the 'help' condition.[88] In each case of course the chairman has the same information available to him concerning the side-effect in question – essentially relying on the vice-president's claim that the program will have the relevant effect – and so the

[88] Corroborating earlier results examining subjects' intuitions as to whether an agent *knowingly* brought about another's death in a 'morally blameworthy' or 'morally neutral' condition (Nadelhoffer, 2006).

asymmetry presents a potential problem for a 'scientific' account of the knowledge concept according to which it identifies some (psychological) natural kind. If 'knowledge' is a natural kind, the criteria for knowing something should be independent of any moral valence placed on the object of that knowledge.[89]

Performance error accounts of the asymmetry, which would preserve the response-independent quality of knowledge as a category, proceed along similar lines to those already seen. It could be argued, for example, that participants are subject to a kind of 'theory contamination': they *would* make symmetrical knowledge attributions in each case, except that they hold an explicit theory that an agent can only be blamed for the consequences of her actions if she knew (or believed, or ought to have believed) that her action would bring them about, such that the intuitive knowledge attribution is modified to maintain this theory. Alternatively, one might present the explanation in terms of 'affective interference': negative affect toward the harming chairman interferes with our otherwise neutral knowledge intuitions so that the attribution is made that seems most expressive of condemnation (similar to a 'conversational pragmatics' explanation whereby the knowledge attribution is simply a way of avoiding the undesired implication that the agent did not know of and therefore *cannot be blamed for* the harmful side-effect).

As with the previous asymmetries, these explanations become a little more stretched when the asymmetry in question is expressed not in terms of moral valence but of norm violation more generally. This is the case when scenarios are used that involve norms contrary to participants' own moral norms, as in the following vignette from (Beebe, 2013), modified from (Knobe, 2007):

> In Nazi Germany, there was a law called the "racial identification law." The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes. The vice-president of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be violating (/fulfilling) the requirements of the racial identification law." The CEO said: "I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!" As soon as the CEO gave the order, the corporation began making the organizational changes. (Beebe, 2013, p. 238)

When asked whether the CEO knew (or believed, or should have believed) that he would be violating the race identification law, participants were significantly more likely to make

[89] Alfano, Beebe and Robinson extend the asymmetry from claims about what the chairman 'knew' to similar claims about what he 'believed' or even what he 'should have believed' on the basis of the evidence (2012).

knowledge attributions in the norm-violating (law-breaking) scenario, even though this involved an effect to which participants would have attributed a positive moral valence.

The conceptual competence explanation of these results is what has been called the 'deliberation' or 'doxastic heuristics' model of knowledge attribution (Alfano, Beebe, & Robinson, 2012; Beebe, 2016). The idea is that agents contemplating norm-violating courses of action are assumed to engage in a higher degree of reflection on the situation, which is reflected in the increased likelihood that participants will agree that they knew, believed or ought to have believed that the norm-violating side-effect would occur. Though Beebe (2016) argues that this explanation retains the 'epistemic purity' of knowledge (i.e. the idea of knowledge as a natural kind) it is clear that in practice what 'counts' as knowledge depends on normative elements of the situation in which it features. A moderately strong source of evidence (for example, the vice-president's statements) counts as knowledge when it warns of a norm violation that an agent proceeds to ignore, but doesn't count as knowledge when it informs of a norm-following outcome that an agent completely disregards.

What the above asymmetries hopefully illustrate, taken together as part of the 'person-as-moralist' picture of folk-psychology and causal reasoning, is how pervasive a role our perceptions of norms, both moral and otherwise, plays in our cognition. The fact that many everyday concepts that feature in our folk-psychology and causal cognition (causation, intentionality, knowledge, and many others) are sensitive to these norms, and to norm-violations in particular, should not strike us as all too surprising. When it comes to explaining these interactions, I have argued that the 'performance error' accounts required to maintain a realist assumption for each of these concepts – such that they can be understood as referring to true natural kinds – are profoundly unsatisfying. It is certainly possible to come up with such explanations, but because they involve writing off a significant number of intuitions as mistaken, the underlying accounts of what the natural kinds in question actually *are* can draw no real empirical support from folk intuitions themselves. The underlying 'facts' with which our supposedly unreliable intuitions are contrasted ultimately rely on a kind of *a priori* that we are expected to accept without more.

Beyond this, the kind of explanations that performance error accounts rely on are in a fundamental sense not properly *explanatory*. Explanations in terms of something like 'motivated cognition' for example – the idea that agents modify their intuitions out of some

desire to maintain certain core beliefs or justify some distinct judgments – are a useful way of describing particular patterns of behaviour but not much more. This is because the kind of mental states in question are always simply postulated, as 'unconscious' beliefs or desires capable of operating on our unconscious cognitive processes, such that the only evidence for them is the very pattern of intuitions that they are supposed to explain. The stipulation of these various deep psychological facts is not only unparsimonious generally, but it only makes sense if we already accept that there is something (some deviation from expected patterns) that is in need of an explanation in the first place – and this is only the case if we have already accepted some *a priori* theory of what the folk-concept *ought* to be picking out.

The alternative response to the empirical results, of which the available 'conceptual competence' accounts of the asymmetries considered above are good examples, is to abandon the realist assumption altogether. In doing so, we are able to make sense of our folk practices as demonstrating an underlying competence with the concepts in question in a way that does much better justice to the empirical results. The abandonment of the realist assumption need not trouble us all that much, as ultimately it may be unrealistic to expect the contrary: that concepts that have developed to pick out features of the world that are relevant *to us* should also pick out features of the world that happen to be natural kinds (as opposed to response-dependent properties dependent in some way on facts *about us*). It is for this reason that, in the next section, I will attempt, drawing on the asymmetries examined above, to provide a 'conceptual competence' account of the empirical observations concerning the folk concept of the deep self.

§4) The Deep Self in Empirical Studies

So far the theoretical accounts of the deep self that I have considered (in Chapter II) have approached the deep self on the basis that it constitutes a real, objective feature of the world – a natural kind – the existence of which is indicated primarily by our deep self intuitions. In this chapter I have so far summarised the results from empirical investigations into other folk concepts and argued for what I have described as a 'conceptual competence' view that entails the rejection of any kind of realist assumption. I now turn to certain more recent empirical studies that more specifically explore the folk concept of the deep self, approaching it as just that – a folk concept – without any of the ontological assumptions or commitments that have characterised philosophical accounts of the deep self in general. It is worth noting that, whilst relatively new for philosophers, this kind of empirical investigation into the deep self as a

psychological construct is already well established within experimental psychology: many studies have shown the deep self construct to not only to feature in ordinary thought but to play an explanatory role in a range of behavioural phenomena.[90] As such, it will be helpful to attempt to draw comparisons between the psychological literature on deep self attributions and the emerging empirical results within philosophy.

(i) The Essential Elements of Identity

The starting point for explorations of the deep self within experimental philosophy is an investigation into the concept of personal identity and those features of persons that are most central to their identity. Philosophers have long been interested in what constitutes an individual's identity, wanting to know what makes in individual who they are, or what kind of changes would mean that they were 'no longer the same person'.[91] Experimental investigations into folk intuitions of identity and identity continuity have involved presenting test participants with scenarios similar to those used as thought experiments by many traditional philosophers: situations in which some aspect of an agent's identity is changed and the question is asked whether the agent is still perceived as 'the same person' as before.[92] Strohminger and Nichols in particular carried out a range of such investigations, coming to the conclusion that *moral* traits are the most relevant to the folk concept of personal identity (2014, 2015). In vignettes ranging from brain transplants and pharmaceutical interventions to the effects of ageing, participants identified personality traits like honesty, trustworthiness, compassion, generosity, etc. (characterised by the authors as 'moral traits') either as more likely to be retained across changes, or as being more important to an individual's identity (such that changes in those traits were more likely to be perceived as fundamentally identity-altering or identity-destroying). This was in contrast with a range of 'non-moral' personality traits – basic cognitive traits, desires and preferences, etc. – so as to lead the authors to conclude that "not all parts of the mind are equally constitutive of the self", and finding "strong and unequivocal support for the

[90] As just one example, experimental psychologists have explored the way in which deep self attributions (and in particular perceived disparities between subjects' 'true selves' and their behaviour) seem to play a role in predicting addictive behaviour and motivation for abstinence (Downey, Rosengren, & Donovan, 2000).

[91] One example is the work of Derek Parfit on personal identity and what he has described as 'Relation R' – the kind of psychological connectedness or similarity that seems to matter to us when thinking about identity and survival.

[92] Whilst the framing of the question is admittedly ambiguous as between numerical and qualitative identity, as none of the scenarios involve duplicates or destruction of one individual and replacement with another it can be assumed that the folk concept they are getting at is that of qualitative identity – whether participants would *treat* individual B as 'the same as' individual A.

essential moral self hypothesis" (Strohminger & Nichols, 2014).[93] Preliminary explanations by the authors put the centrality of moral traits down to either (i) a belief that moral traits are central to what it is to be human (such that changes in these traits are perceived as identity-destroying) or (ii) the fact that moral traits are likely to have a greater social utility, for the evaluation of social partners.[94]

What emerges from later studies is a kind of 'identity asymmetry' similar to the range of asymmetries considered in the previous section. What is revealed through the investigation of folk understandings of personality change is that, where an agent undergoes a change that is evaluated as 'positive' (such that some moral trait is improved) the change is typically explained in terms of 'self-discovery', suggesting that participants saw these changes as fundamentally continuous with who an agent always (really) was all along (Bench, Schlegel, Davis, & Vess, 2015). Where, on the other hand, an agent undergoes a change that is evaluated as 'negative', the change in question is more likely to be understood as 'identity-destroying', in the sense that the agent afterwards is no longer really herself (Tobia, 2015, 2016). As with many of the asymmetries explored earlier, it seems likely that rather than mapping directly onto the 'moral valence' of personality changes, people's identity intuitions are sensitive to changes that are 'norm-following' or 'norm-violating' in a more general sense.

One source of evidence for the 'norm-violation' interpretation is a study by Heiphetz, Strohminger and Young (2017) which found that, in addition to the general effect of 'direction of change' (positive trait to negative or the reverse) changes in 'controversial' moral beliefs were perceived as less identity-altering than changes in widely held moral beliefs. One way of explaining this is that whilst any change to a widely held moral belief would be seriously norm-violating – in the sense that it is counter to both prescriptive (moral) norms (what agents *ought* to do) and descriptive norms (what agents generally *actually do*)[95] – the potential violation of a

[93] The authors point out that these results appear to conflict with earlier results concerning personal identity that have highlighted the importance of traits like memories, but argue that this is explained by the focus of earlier studies on the first-person experience of identity (where without memories we would no doubt feel seriously 'unmoored') and the new focus on the third-person experience of identity (from which perspective the loss of another person's memories is not nearly so significant as long as other traits – moral traits – remain intact).

[94] There is some empirical support for an explanation along these lines from a study by Heiphetz, Strohminger and Young (2017) showing that the relationship between different kinds of change and the extent to which they were perceived as identity-altering was mediated by participants' predictions about how much the changes in question would affect their relationships with the agent. For example, whilst changes in widely held basic moral beliefs were generally perceived as seriously affecting identity, changes in controversial moral beliefs on topics like abortion were perceived as less significant because they were less strongly linked with community relationships.

[95] I recognise that many philosophers may not understand the 'descriptive norms' that I refer to here as norms at all. The reason that I refer to both as simply different kinds of norms is that they are two different ways in which

prescriptive norm that a change in some controversial moral belief would represent is moderated by the fact that the resulting belief is still 'normal' in the descriptive sense that many other people share it. So, for example, whilst from the perspective of a pro-choice campaigner another agent's change of moral belief from pro-choice to pro-life might seriously violate moral norms, this is likely to be perceived as less identity-altering than a change in some more widely held moral belief because the resulting belief in the first case is still not all that uncommon.[96]

(ii) Asymmetrical Deep Self Attributions

None of the experiments considered so far has addressed the deep self concept we are interested in directly, though they may have touched upon it indirectly via intuitions of identity and continuity. The general hypothesis concerning the folk concept of identity is that, when thinking about *who somebody is*, we have a tendency to assume that others are 'good' (or in some sense 'norm-following') in a way that produces a kind of asymmetry in our identity intuitions. When we begin to investigate deep self attributions directly, the same kind of asymmetry is once again revealed. Newman, Bloom and Knobe (2014) began a study by presenting participants with vignettes in which an agent undergoes some kind of (normatively weighted) change and asking them what part of the agent was responsible for the change: her deep self ("the deepest, most essential aspects of her being") her surface self ("the things that she learned from society or others")[97] or something else, as well as a 9-point scale for indicating the extent to which the new behaviour represented the agent's true self. Participants were significantly more likely to attribute 'positive' changes to an agent's deep self and negative

me might think of something like 'normal' behaviour. What is important to note is that from a folk-psychological perspective there is significant overlap between the two concepts: ideas of what is prescriptively normal for a given category influence judgments about what is descriptively average for that category (Bear & Knobe, 2016).

[96] Another potential source of evidence for the 'norm-violation' interpretation comes from the findings of Molouki and Bartels (2017) that the extent of the 'direction of change' asymmetry is moderated by people's expectations. Changes to personality traits understood in terms of 'deterioration' (as opposed to improvement) were less identity-altering when participants expected the deterioration in question. Though Molouki and Bartels' results concerned first-person perceptions of change and continuity, it seems just as plausible that the same might be true of our identity intuitions concerning others. We might find deterioration in certain moral traits in an agent to be less identity-destroying if it were part of some normal (expected) developmental process – consider, for example, the personality changes that agents commonly undergo during their teenage years. Similarly, we might find such deteriorations less identity-destroying if they involve a return to some kind of 'average' baseline: even if we admire the moral traits of, for example, a person who has been a committed vegan, we are likely to perceive the loss of these traits as something like the 'end of a phase' and so ultimately as a return to that person's true, underlying identity.

[97] Admittedly the cashing out of 'surface self' in terms of things that an agent "*learned from society or others*" is not ideal, as it inserts a particular theoretical belief about the nature of the deep self into the experiment, but as long as the contrast with the "*deepest, most essential aspects of her being*" is in place it seems likely that the items elicited the appropriate 'internal'/'external' distinction that deep self intuitions involve.

changes to an agent's surface self – giving us the basic form of a normative asymmetry in our deep self intuitions.

In a second study they confirmed the normative asymmetry by using items for which the valence would change depending on the normative orientation of the participants: behavioural changes included from 'unpatriotic' to 'patriotic', 'promiscuous' to 'monogamous', 'atheist' to 'believer', 'homosexual' to 'heterosexual' and other items designed to elicit positive or negative assessments from conservative or liberal participants. Participants (who indicated their political orientation as liberal or conservative) then indicated their agreement with the statement (for each vignette) that there had always been something deep within the agent, calling her to behave in the new way. Results showed conservatives far more likely to attribute behavioural changes in the 'good for conservatives' vignettes to something deep within the agents and the same for liberals with respect to the 'good for liberals' vignettes.

Finally, the authors examined the interaction between the observed normative asymmetry and a belief/feeling distinction in the presentation of different motivations. Following the previous study, the vignette was designed to divide liberals and conservatives in terms of their normative assessments of the two competing behaviours: the agent in question was either an evangelical Christian who *believes* that homosexuality is wrong but has conflicting feelings (he is attracted to other men), or a secular humanist who *believes* that homosexuality is perfectly acceptable, but has similarly conflicted feelings (he is repulsed by the thought of same-sex relationships). When asked to attribute either the belief or the feeling to the agent's deep self, though there was a slight main effect of mental state type (participants were, in general, likely to view 'feelings' as more indicative of an agent's true self)[98] this was still subject to the much more powerful effect of normative valence. Essentially, participants were "more likely to regard [the agent's] psychological states as part of his true self when those states fit with their own values" (Newman et al., 2014). What these results suggest, in a very preliminary way, is that people don't have any particularly strong convictions or substantive theory about the *kind* of mental state that makes up the deep self (the question with which the philosophical accounts of the deep self considered in Chapter II were primarily occupied) but are instead willing to attribute to the deep self whatever mental states most neatly coincide with their own values.

[98] The belief/feeling distinction is obviously not a very good representation of the difference between Frankfurtian (cares-based) and Watsonian (commitments-based) conceptions of the deep self, but the fact that participant were somewhat more likely to attribute the 'feelings' in question to the agent's deep self might be consistent with the movement within both traditions away from a conception of the deep self in terms of conscious (propositional) belief and towards less precise mental states over which an agent is understood to have less voluntary control.

There are a number of ways in which the emerging asymmetry might be cashed out: as a tendency to attribute 'good', or perhaps more simply 'norm-following' essences to other agents, or as reflecting an assumption that others are somehow, deep down, like us and share our values. However it is cashed out though, the asymmetry in question represents a serious problem for deep self realists, to which they might respond in one of two ways. The first would be to accept that people generally *do* have 'good' deep selves, but cashing out this 'good' in terms of any particular normative framework still leaves deep self realists faced with the difficult task of explaining why agents within different normative frameworks make different deep self attributions. If the content of an agent's deep self represents an objective (response-independent) reality about that agent's psychology, then the normative standpoint of any given observer can have no bearing on it, and yet we as observers seem unable to separate our deep self intuitions from our normative evaluations. The second way in which the deep self realist might respond is to deny any general claim about the 'goodness' of deep selves and describe the pattern of intuitions as a kind of 'positivity bias'. In much the same way that people's tendency to overestimate their own knowledge and abilities doesn't show that there isn't an objective fact of the matter, the existence of a simple positivity bias in deep self attributions need not worry the deep self realist. The problem is that simple positivity biases can be represented on a single linear scale such that if every agent corrected their positivity bias each of their assessments would converge on reality. When it comes to moral assessments within completely different normative frameworks, there is simply no guarantee that a 'correction' of each agent's positivity bias would lead to all assessments converging on the same point.

Regardless of its relationship to a possible 'objective reality' about deep selves, the pattern of deep self intuitions has been described by empirical investigators as 'deep self optimism': a "general tendency to conclude that deep inside every individual there is a "true self" calling him or her to behave in ways that are morally virtuous" (Newman et al., 2015). The kind of optimism in question, if it can properly be described as optimism, appears to be extremely robust, as the deep self asymmetry has been confirmed across cultures and even across differences in levels of reported misanthropy (De Freitas, Sarkissian, et al., 2017). The robustness of these results suggests that the source of the deep self asymmetry is not straightforwardly attributable to some kind of (even unconscious) 'belief' in the goodness of others, the strength of which might plausibly vary between individuals and across cultures, but likely in some relatively fundamental cognitive mechanism. So far theorists have considered the phenomenon of psychological essentialism (or 'teleological thinking') to be a potential

source of the deep self asymmetry (De Freitas, Cikara, Grossmann, & Schlegel, 2017; De Freitas, Sarkissian, et al., 2017)[99] but the explanation is still far from complete. Though beliefs in the existence of deep selves might display many of the features of psychological essentialist thinking, nothing in our current understanding of psychological essentialism explains *why* the essences attributed should be morally good.

(iii) Attribution Studies

It is worth pausing at this stage to consider a set of studies from experimental psychology referred to as 'attribution studies': investigations into the way in which people explain behaviour through attributions of causal influence. For a long time the results from such studies had been interpreted as revealing a 'self/other' or 'actor/observer' asymmetry: individuals seemed to explain behaviour experienced from the 'actor' perspective (in the first person) in terms of situational influences, but when placed in the position of an observer (in the third person) would explain the same behaviours in terms of 'behavioural dispositions' of the agents in question.[100] This apparent asymmetry was initially explained by contrasting people's intimate awareness, as actors, of the features of a situation to which they are reacting, with the tendency of observers to attend most to the behaviour of other agents. It represented the standard interpretation of attribution studies until a meta-analysis by Malle (2005, 2006) revealed that it holds only under relatively specific conditions – most significantly, only in relation to *negative* events. When the normative valence is reversed, the opposite effect appears: participants become more likely to attribute other agents' behaviour to elements of the situation. As Newman and colleagues summarised the results: "for behaviours performed by themselves, participants tend to attribute the good to the person and the bad to the situation, but for behaviours performed by others they tend to attribute the good to the situation and the bad to the person" (2014).

This normative asymmetry in attribution studies appears to present a challenge to the 'deep self optimism' hypothesis. It seems odd that people should on the one hand overwhelmingly attribute 'virtuous' essences to other agents in explaining their behaviour yet at the same time tend to see other agents' bad behaviour as attributable to them and good

[99] The general phenomenon known as psychological essentialism is by now extremely well-documented. The general idea is that humans have a cognitive tendency to think of certain categories as having underlying shared 'essences' that account for their common features and persist across superficial changes – for an overview see (Gelman, 2003; Gelman & Medin, 1993; Haslam, Bastian, & Bissett, 2004; Medin & Ortony, 1989).

[100] For a relatively early example, see (Kelley, 1967, 1973).

behaviour as attributable simply to features of their situations. My proposal is to explain the normative asymmetry in attribution studies – and ultimately the asymmetry in deep self attributions – in the same terms as the principal asymmetries considered in the previous section: in terms of norms and norm violation and the central role that they play in our causal cognition and folk psychology. We should start by establishing what exactly our 'attributions' in attribution studies are. Essentially they are answers to the question 'why did the agent act in the way she did?' – or, more specifically, 'what *caused* her to act in the way she did?', where potential explanations might attribute causal influence either to features of the situation or to features of the agent herself. In this way the task activates people's causal cognition – which, as we know from the previous section, works by directing our attention towards those features of a situation that are likely to be most explanatorily useful. We then need to explain, with respect to our intuitions as observers, why situational factors are more explanatorily useful when it comes to norm-following (positive) behaviour and agential factors are more explanatorily useful when it comes to norm-violating (negative) behaviour.[101]

To see how this might be the case, consider first the following scenario: a person, walking by a pond, sees a child who has fallen in and is drowning and leaps into the pond to rescue it (doing, we might imagine, what anyone *would* do in such a situation). When it comes to explaining *why* somebody jumped in a pond, referring to their good moral qualities – perhaps some general disposition to help others – is not particularly explanatory: lots of people have such moral qualities and never jump into ponds in their lives. A better explanation would refer to the features of the particular situation with which this person (with his relatively normal moral qualities) was faced: there was a child who *needed help*. Consider the opposite scenario: a person walking by a pond sees a child drowning and *doesn't* leap in to rescue it. One might suggest, for example, that the person does not leap into the pond because it is a cold day and she does not wish to get wet – a situational factor that is generally quite a good explanation for why people don't jump into ponds. But it would be unhelpful to refer to these features of the situation to explain this particular behaviour because, we imagine, most people in the situation *would* leap into the pond. The general behavioural predictions that one would make on that basis – that chilly weather causes people not to jump into ponds to save children – would likely

[101] 'Positive' and 'negative' are used here to reflect the ordinary way in which norm-following and norm-violating behaviours have been represented. It remains possible, of course, that 'positive' behaviours (in the sense of *morally good* behaviours) may violate descriptive norms of what people ordinarily do. This kind of supererogatory behaviour has been underexplored in the empirical literature, which has focused on blaming reactions over praising, but would provide an interesting opportunity to test the hypotheses presented here.

be wrong. In explaining why *this* person acted in a way that others wouldn't, it is most explanatorily useful to refer to features of the agent herself. Because the agent has behaved in a norm-violating way – responding to the situation in a way that other agents would not – the features in question will necessarily be features that differentiate her from others or else they will not be explanatory.[102] Norm-following behaviour, on the other hand, is best explained not by reference to particular features of the agent in question but by reference to the norms themselves (for which situational features are a kind of proxy). The only reason that features of a situation, (such as a child needing help) are explanatory at all in such cases is because we understand that, given those features, the *normal* thing to do is to jump into the pond.[103]

The more basic explanation of the attribution asymmetry, which suggests that as folk-psychologists we have a tendency to attribute the 'bad' of other people's behaviour to them and the 'good' to situational factors, obscures two things. Firstly, the descriptions of behaviour as 'good' or 'bad' in the relevant scenarios obscured the fact that invariably the distinction was between norm-following and norm-violating behaviour, where prescriptive and descriptive norms coincided. Secondly, a focus on the tendency to attribute bad actions to agential features and good actions to situational features obscures the general assumption being made that other agents are basically norm-following. The reason why norm-following behaviour is typically explained by reference to the features of the situation is that is *goes without saying* that people generally act according to the norms of such a situation. On the other hand, our visible (or conscious) attributions of causal influence to other agents only ever involve norm-violating behaviour because it is only norm-violating behaviour that *demands* such an explanation. This is essentially consistent with the studies concerning the folk concept of causation referred to earlier: greater causal significance is likely to be attributed to norm-violating elements of a situation because it is those elements that produce the most useful causal models.

§5) Explaining the Deep Self Asymmetry

[102] This need not involve 'deep self' attributions at all. Whilst the behaviour in question might be explained by some kind of moral failing of the agent in question, it could equally well be explained, for example, by his not being able to swim. The point is simply that, when faced with behaviours that are unusual (norm-violating) in a given situation, it is most explanatorily useful to refer to agential rather than situational features.

[103] The extension of this hypothesis, which has yet to be addressed by empirical studies, is that when faced with instances of supererogatory behaviour – unusual but highly admirable actions like running into a burning building to save someone – subjects will show more ambivalent responses, likely referring both to features of the situation that invoke the relevant moral norm ('somebody needed saving') and features of the agent that explain her violation of the descriptive norm ('she was particularly brave').

Returning to the deep self intuitions with which we are primarily concerned – including the observed tendency to attribute morally good deep selves to other agents – we can begin to see how it might be explained in similar terms to the attribution asymmetry considered above. It essentially involves the same mechanism as the attribution asymmetry applied not to a distinction between agential and situational factors but to that between surface selves and deep selves. The main difference is in the kind of exercise with which participants are presented: participants in attribution studies are ordinarily presented with a single instance of behaviour and asked whether it is attributable to elements of the situation or some feature of the agent; participants in the recent deep self attribution studies have been presented with two 'versions' of an agent – represented either as before and after a change or as an internal conflict – and asked to attribute one of them to the agent's deep self and the other to her surface self. Whilst attribution studies have a tendency to elicit 'negative' agential attributions (attributing norm-violating behaviour to features of an agent) and 'positive' (default) agential attributions have been obscured, deep self attribution studies always allow subjects' default deep self attributions (of norm-following deep selves) to manifest themselves. If subjects make an underlying assumption that agents' deep selves call them to act in norm-following ways, then they always have an opportunity to 'choose' the deep self attribution that confirms this.

Of course the idea that norm-following deep self attributions represent a sort of default assumption certainly doesn't mean that we don't regularly make negative attributions, even if this has not been captured by the specific experimental designs in question. When faced with bad (norm-violating) behaviour we very regularly *do* resort to such attributions – just as our attributions of intentionality or our choice of causal explanation do represent underlying *negative* deep self attributions. As Sripada notes with respect to the intentionality asymmetry, the judgment that the Chairman *intentionally* harmed the environment maps onto an underlying judgment that this negative trait (disregard for the environment) is attributable to the Chairman's deep self (2010). There are therefore three questions that still need to be answered: (i) why it is that we make deep self ('essence') attributions at all; (ii) why it is that our 'default' essence ascriptions tend to involve morally virtuous characteristics; (iii) what is it that causes us to deviate from these 'default' deep self ascriptions in particular cases in the way that we do?

Starting with the first question, a useful framework within which to consider role of deep self attributions in our folk-psychology (or indeed the more general phenomenon of psychological essentialism of which it appears to be an instance) is the 'model complexity framework' recently put forward by Nichols (2017). Essentialist explanations represent the

probabilistic relationships between a range of features and a single postulated 'common cause', whereas other models might for example represent the full range of probabilistic relationships between each feature. Nichols argues that the use of 'common cause' models is, given a few assumptions, computationally rational – meaning, essentially, that it is a good way of using incomplete data and Bayesian learning to make inferences about unknowns. Considerations of model complexity favour common cause models because they have fewer parameters, making the essentialist model "the least flexible model and the one that should be preferred in model selection, all else being equal" (Nichols, 2017). Put in terms of our folk-psychology, the availability of (essentialist) deep self attributions as a way of postulating common causes for a wide range of observable behaviours is a computationally rational way to model human behaviour in order learn and make predictions about the future.

As to why there is a general tendency to attribute 'virtuous' essences to agents, we must start by recognising that the normative asymmetry in question does not only apply to our folk psychology. De Freitas and colleagues (2016) have shown the asymmetry to apply to essence ascriptions for a range of entities beyond human agents, demonstrating that normative beliefs play a similar role in our beliefs about identity continuity for entities ranging from nations and universities to bands and academic papers. Just as in the human identity continuity studies, participants were presented with a vignette involving some entity (such as a nation) going through either a positive or a negative change (in the case of the nation, becoming either more tolerant or more discriminatory) and asked whether they thought that it was still the same entity that it was before the changes or that the essence of the original entity had now ceased to exist. Participants reliably perceived the positive change as continuous with the entity's 'essence' and the negative change as involving identity-loss. The reason is most likely that (in the absence of any other information) participants were "less likely to say the entity was the same when it possessed negative properties simply because it was no longer seen as satisfying the definition of the category to which it previously belonged" (De Freitas et al., 2016).[104] This is consistent with other studies indicating that the construction of category essences – of what is 'normal' for a given class or kind – involves a combination of both descriptive (statistical) and

[104] Initial results suggest that this holds true for entities with a morally 'negative' essence as well, as shown in De Freitas and colleagues' final study (De Freitas, Tobia, Newman, & Knobe, 2016). Participants were more likely to say that a 'bad' entity had lost its identity when it underwent positive as opposed to negative change, although the study did not explicitly address negative *category* essences. It remains a possibility that when faced with a negative category essence people may nonetheless attribute positive change to an underlying positive *individual* essence, with the resulting entity having lost its category identity but not its individual essence.

prescriptive (moral/evaluative) norms (Bear & Knobe, 2016).[105] Our essence ascriptions to other agents are likely to take as their starting point the assumption that those individuals will conform to the norms of the category to which they belong – that is, that they will act in 'norm-following' ways for human agents.

The third question – concerning the reasons why we deviate from these default assumptions when we do – is the most important because of the obvious link between the norm-violating behaviour in question and the moral responsibility judgments in which we are interested. Why is it the case that some instances of norm-violating behaviour are held to be non-attributable while others are held to be attributable? Why do we sometimes retain our default essence ascriptions (explaining norm-violating behaviour by reference, for example, to features of an agent's environment) whilst at other times explaining norm-violating behaviour by ascribing a norm-violating essence to the agent? Consider an instance of norm-violating behaviour considered earlier: an agent, walking past a pond, fails to jump in to save a drowning child. What we first need to recognise is that we approach a situation like this with a background model of 'normal' human behaviour (influenced by a range of factors, including moral norms and salient social narratives) that provides us with expectations (predictions) about how an agent is likely to behave in that situation. When the agent doesn't act to save the child, those expectations are violated. Violations of our expectations require some kind of explanation, but it need not involve a deep self attribution. Whilst we might explain the behaviour in question by assuming that the agent in question doesn't care about saving children, we could equally explain it by assuming that she can't swim. So what determines which of the two explanations we intuitively decide on?

One potential explanation – that will no doubt be preferred by deep self realists – involves some kind of 'motivated cognition'. It might involve the claim that people make deep self attributions when they want to blame the agents in question, so as to justify their blaming responses.[106] An alternative explanation could involve the claim that people make deep self attributions when it is *useful* to blame the agents in question, relying on some evolutionary story about how we came to be sensitive to the features of such situations. On this model, our

[105] The authors concluded that, in coming up with a concept of 'normality' for a given category, "*people's representations of what is normal were […] influenced both by what they believed to be descriptively average and what they believed to be prescriptively ideal*" (Bear & Knobe, 2016).

[106] The motivated deep self judgment does not immediately translate to blame, as it is only the attributability step of the attributability/accountability pair, but if an agent is motivated to hold another accountable, this would create a motivated cognition-style explanation for the attributability judgment in question.

tendency to blame certain norm-violating actions or agents is due to the fact that doing so tends to minimise the norm-violating behaviour in question. Our tendency *not* to attribute blame for certain kinds of norm-violations would be due to the fact that such blaming responses would have no useful effect. For example, if blaming people for the range of behaviours that we understand to be 'compulsive' – including things like addictive behaviour – has no useful effect, we should expect people's deep self attributions to reflect this such that we no longer blame the agents in question. For a number of reasons I am disinclined to adopt this kind of model. First of all, like the various performance error accounts discussed above, the principal evidence for the postulated mental states (the desire to blame, or the belief that it is useful to blame) is the pattern of behaviour that they are supposed to explain. Secondly, such a model relies on a conflation of attributability judgments and accountability judgments, assuming that our attributability judgments are always followed by expressive blame or punishment. Finally, as I will argue below, there may be a viable alternative model that is capable of explaining the pattern of deep self attribution by reference only to already established models of human cognition and folk-psychology.

What I have in mind is the following: whether or not the 'difference-maker' required to explain a given instance of norm-violating behaviour involves a deep self attribution or not depends on which of the available explanations is the most parsimonious and requires the least modification of the observer's underlying model of human behaviour. One potential way of understanding this is as a kind of 'conservatism' when faced with competing explanatory models: the explanation will be preferred which requires the least modification of the existing model. What the 'conservative' option turns out to be will depend, of course, both on the underlying model and on the available (competing) explanations, and as a result might vary from person to person.

Let us return to the example of the agent who violates our expectations by not jumping into the lake to save a drowning child. The deep self explanation of this behaviour is, against the background of my underlying model of human behaviour, not particularly parsimonious: it requires me to assume something (an agent not wanting to save a drowning child) to which my underlying model had assigned an extremely low prior probability. In other words, everything I think I know about human behaviour tells me that it is extremely unlikely that someone would not jump into a pond to save the child. The advent of this (on my initial model) highly unlikely fact requires a correspondingly significant modification of the prior probability assigned to agents not caring about children. The alternative explanation – the one involving the assumption

that the agent in question can't swim – fares a little better. The assumption in question is likely relatively parsimonious, given that my underlying model of human behaviour probably assigns a decent prior probability to the possibility of an agent not knowing how to swim.[107] From that point everything else in the scenario proceeds as expected for an agent who doesn't know how to swim: people who can't swim are highly *unlikely* to jump into ponds, even to try and save drowning children, because they themselves might drown in the process.

There are a few important features of this model that are central to explaining the variability in people's deep self intuitions. The first is the fact that our deep self intuitions are never simply a way of explaining individual instances of behaviour, or even an individual agent's pattern of behaviour. The 'goal' of our deep self attributions is always to provide stable and predictively useful models of human behaviour, and so the 'stimulus' out of which our folk-psychology is trying to make some kind of sense is always a whole spectrum of human behaviour. The second is that deep self attributions, as a particular kind of causal explanation, are always in competition with other kinds of causal explanations to provide an explanatorily useful model of human behaviours with which we are confronted. Addictive behaviour presents a useful illustration of this: it could be explained by positing some underlying trait of a certain class of people (something amounting to a moral failing, perhaps) disposing them to act in the relevant norm-violating way; but the phenomenon of addictive behaviour could equally be explained by positing an alternative common cause, located not in the agents themselves but in the situations in which addictive behaviour is manifested. The latter kind of explanation is what we are engaging in when we talk about something like chemical addiction, postulating a situational feature (the chemical properties of addictive drugs) to which all of the agents in question have been exposed and which causes the behaviour in question.[108] A final thing to note is the way in which a different underlying model of 'normal' human behaviour can lead to different deep self intuitions. To the extent that subjects with different moral views from one another (different prescriptive norms) have correspondingly different conceptions of what 'normal' behaviour is, the underlying causal model that they bring to any particular situation

[107] The argument is not that people will intuitively excuse the agent for not jumping into the pond, but rather an illustration of how competing explanations might work. No doubt many people's intuitive assumption when faced with such a scenario is that the agent in question is simply *bad* – perhaps because moral (deep self) explanations are a particularly common category or in some other way salient. If in some way prompted, however, to the possibility that the agent might, for example, not know how to swim, my contention is that this latter explanation would generally prove easier to accept.

[108] In this sense deep self attributions are only one kind of essentialist explanation. Explanations that postulate underlying properties in other entities, such as the disposition to cause a certain kind of behaviour in human agents, equally involve attributing causally potent 'essences' in order to explain observable features.

will be fundamentally different. The reason why the Mormon is more likely to attribute a conflicted agent's abstinence commitments to her deep self is the same reason that another subject might be likely to attribute her sexual desires to her deep self: in each case the deep self attribution in question is the most parsimonious and requires the least modification of the underlying predictive model. Moreover, the causal model entailed by each will probably also turn out, in the context of each agent's very different life, to be a more useful source of behavioural predictions for the kinds of behaviours and agents that each is more likely to encounter. The greater the modifications we are required to make to our underlying causal models in order to accommodate norm-violating behaviours, the less predictively useful those models will be in the majority of cases. For the Mormon to change her underlying causal model of human behaviour any more than necessary in order to explain the behaviour of a non-community member will make that model less useful for predicting the norm-following behaviour of the kind of people she might be more likely to interact with on a regular basis.

One potential objection to the model that I have presented so far is that we sometimes seem to blame people – meaning that we attribute to them some norm-violating deep self – for behaving in ways that conform to our expectations of totally standard human behaviour. One example of this might be the morally upright public official in a corrupt country: although almost everyone he interacts with offers him bribes, to the point that he expects this behaviour of those around him, he may nonetheless blame them for doing so.[109] This would potentially undermine the account of deep self attributions as a certain kind of explanation required to explain unexpected behaviour. I would argue however that in cases like this there is still a sense in which the behaviour in question, despite being consciously understood as descriptively normal, violates a certain expectation – perhaps even a subconscious expectation. The 'moral reformer' (who condemns the 'normal' behaviour of those around her) still remains hopeful that people will act according to what she sees as moral norms, which perhaps manifests itself in a kind of subconscious expectation. Supporting this account is the fact that our *own* behaviour likely plays an important role within our folk-psychological model of human behaviour, as one important way in which we come up with predictions of how others will behave in novel situations is by imagining our own behaviour in such situations. It is thus possible that the agent

[109] Of course the model I have proposed does not rule out the possibility that in some cases people's underlying model of human behaviour *will* involve a default assumption of morally 'bad' essences, and in these cases morally 'bad' behaviour will not require any kind of 'difference-maker' explanation because the default deep self attribution – with the moral condemnation that it entails – will suffice. No modification is needed to the underlying causal model, because the behaviour does not violate any descriptive norms, but the moral consequences flow from the violation of the prescriptive norm alone.

who believes that *she* (unlike others in her community) would behave morally in some situation approaches that situation with the unconscious expectation that others might do the same.[110]

The interaction between prescriptive norms and ideas about what is descriptively normal might seem to undermine the clarity of the model that I am proposing, but my response to this is simply that it reflects the way that we conceive of normality. Whilst the two might often seem to overlap, interesting cases can also be considered in which moral norms and statistical norms are in conflict with one another. Consider the practice of eating meat, seen from the perspective of a vegetarian: on the one hand, the practice constitutes a straightforward violation of a moral norm to the vegetarian, but on the other hand it is also an extremely widespread, 'normal' practice. So what does my account predict about the vegetarian's attributability judgments in this case? The answer is not clear, and would depend on whichever of the two conflicting norms is a more salient feature of the mental model of normal behaviour of the subject making the attributability judgment. A focus on the moral norm might lead to the behaviour in question being attributed to some underlying (norm-violating) disposition in the meat-eater – perhaps some fundamental lack of regard or concern for living things – whereas a focus on the statistical norm might lead to the (norm-following) behaviour being attributed to some situational factor – perhaps explained in terms of the meat-eater's upbringing or socialisation or cultural norms. The fact that my account does not specify which causal explanation will be chosen, far from being a weakness, arguably accurately represents a real ambiguity in terms of how vegetarians appraise the habits of meat-eaters: some condemn meat-eaters for a perceived moral failing while others perceive the behaviour of meat-eaters as a product of their acculturation in a meat-eating society for which they are not really responsible.[111]

Where all this leaves us with respect to sceptical concerns about the folk concept of the deep self is with a viable 'conceptual competence' explanation of people's attributability intuitions. Deep self talk reflects an underlying rational cognitive mechanism – a fundamental

[110] It is even possible that agents who *themselves* participate in some ubiquitous wrong will still respond in the same way, based simply on the actor/observer asymmetry (the tendency of the agent to ascribe her own bad behaviour to features of the situation and that of others to features of the person). It is thus plausible that an unfaithful partner, for example, might still experience some violation of her expectations in discovering that her partner, too, was unfaithful, if she has successfully rationalized her own behaviour as not deriving from any kind of moral deficiency but retains the view that the behaviour in general is do derived.

[111] It seems plausible that the same would be true in the opposite scenario, where behaviour conforming to moral norms seems to violate statistical norms of average behaviour, as in cases of supererogatory behaviour. Whilst the focus of most analyses of deep self attribution so far has been on attributing blameworthy motivations, there is no doubt just as much to be said about the attribution of praiseworthy motivations. An investigation into folk intuitions surrounding supererogatory behaviour would provide an interesting opportunity to test the proposed account of deep self attributions.

part of our folk-psychology – serving to produce and edit the causal models of human behaviour that allow us to make important predictions. To the extent that a particular deep self attribution represents the explanation of some behaviour in terms of a 'causal essence' located in a particular agent, it becomes impossible to say whether such an attribution is 'correct' or not in any objective sense – it is only rational or irrational against the background of a particular cognitive model. There is nothing in this account of the cognitive mechanisms behind our deep self intuitions that 'disproves' any ontological claims about the nature of the deep self as a natural kind. A deep self realist could easily accept the cognitive explanation that I have proposed for our deep self attributions and nonetheless insist that there exists a natural kind of psychological features properly described as the deep self, and therefore a response-independent, objective fact of the matter about the content of any particular agent's deep self. This kind of 'performance error' account remains open, and with it the conclusion that a significant amount of our deep self intuitions are mistaken. But why would we take it? As we saw in Chapter II, there are no convincing proposals for what the natural kind described by our deep self concept and (unreliably) identified by our deep self intuitions might be. Moreover, the full range of our deep self intuitions is already accounted for by the cognitive mechanism that I have described – we don't *need* to postulate, in addition, that there is a corresponding natural kind in the world. But I doubt that this argument will do very much to change the deep self realist's mind, committed as he already is to some *a priori* theory of what the folk concept *ought* to be picking out. This is why, in the next chapter, I consider some of the practical implications of a few different ways of thinking about deep self properties as response-dependent properties, arguing that either an error-theory approach to our deep self intuitions or a form of relativism is preferable to the realist position.

Chapter IV: The Anti-Realist Alternatives

§1) Introduction

So far this thesis has been primarily concerned with raising doubts about the realist assumption in our approach to the deep self concept. In Chapter II I argued that none of the existing theoretical accounts of the deep self provide us with a viable response-independent property capable of characterising a natural psychological kind picked out by our deep self intuitions. Chapter III looked directly at the intuitions themselves, arguing that both the variability in deep self intuitions and the range of factors to which they are apparently sensitive all count against the plausibility of a realist account. I also proposed a few different explanations of the cognitive processes behind these intuitions, arguing specifically for an explanation in terms of model-theoretic properties like parsimony as the determining factors behind our deep self attributions. I have presented these arguments in order to convince the reader that the realist position – that some deep self attributions correctly represent objective psychological facts about the agents towards whom they are directed whilst other deep self attributions get it wrong – is highly implausible. But I also hope to have shown that the kind of criteria proposed by the realist for identifying 'correct' deep self attributions – whether it be as 'cares' or evaluative 'commitments' – have no hope of resolving any kind of disagreement between people who make different deep self attributions: they are simply too vague and flexible to provide a definitive answer. The goal of this chapter, in sketching a range of meta-ethical frameworks that one could potentially adopt as alternatives to this realist position, will be to present some tentative hypotheses about how this kind of disagreement might instead be dealt with.

I will begin by outlining a few different ways in which we could choose to frame our deep self attributions: an expressivist account, a relativist account and an error theory of deep self talk. I will also compare two different practical approaches open to the error-theorist: a fictionalist approach to our deep self talk and an eliminativist approach. I will then work through a single case study of contested deep self intuitions in some depth – the case of drug use and addictive behaviour – in order to make an argument in favour of the kind of meta-ethical approach that I think provides us with the most useful framework for dealing with this kind of disagreement. The way this argument will be framed will be in terms of a social need: social policy requires a particular set of deep self attributions for its justification, and so it is obviously practically useful to be able to achieve some kind of consensus. Whilst I will not present any

particular criteria for determining which set of deep self attributions is 'optimal' (I leave to philosophers with more defined ethical commitments , or at least those with a model for liberal decision-making) I will suggest the framework that I think has the greatest chance, given any particular beliefs about what is 'optimal', of reaching the required consensus – that is, of changing people's minds.

§2) The Meta-Ethical Possibilities

I begin with a meta-ethical approach already alluded to earlier in Chapter III: an expressivist understanding of deep self attributions. The expressivist claim is a claim about the kind of thing that our deep self attributions amount to: essentially, that they are not even trying to be descriptive or fact-stating but simply expressive of a particular kind of evaluative attitude. One way of conceptualising this is by analogy to aesthetic judgments: we might understand the claim that some object is 'beautiful' as simply expressing something like 'Hooray for this object!' – it might not strike others in a pleasing way, but my claim is simply expressive of my own aesthetic sensibility. This kind of approach would lead us to the conclusion that there are no real 'disagreements' about deep self attributions, only expressions of different deep self 'sensibilities' – attitudes, rather than propositions, that don't have any truth conditions, not even relative ones. Deep self attributions could potentially be reduced to the following: a 'bad' deep self attribution expresses something like 'Boo to this behaviour and boo to the perpetrator!' and a 'good' deep self attribution to 'Hooray to this behaviour and hooray to the perpetrator!'.

This kind of account does not necessarily map neatly onto the model-theoretic explanation of the cognitive mechanisms behind deep self attributions that I have proposed. On that particular explanation, the 'reactive attitudes' that constitute our approval or disapproval of another agent are subsequent to and dependent on a deep self attribution that is necessarily explanatorily prior. An expressivist approach might be appealing, on the other hand, to those inclined to accept something like the 'motivated cognition' explanation of our deep self intuitions alluded to in the previous chapter. According to that explanation, whether or not we opt for a deep self explanation over any other kind of explanation when faced with norm-violating behaviour is determined by our desire to attribute praise or blame for the behaviour in question. If we were to accept that kind of explanation about the cognitive mechanisms behind our deep self intuitions it would be a small step to characterise our deep self attributions as expressive of something like '*Praise (/Blame) the perpetrator!*'. Whilst there are various ways in which a set of 'optimal' deep self attributions might be cashed out by the expressivist –

different stories about the kind of deep self sensibility that is most conducive to some description of the common good – the expressivist framework doesn't necessarily give us the tools to get there. This is because an argument about the kind of deep self sensibility that is conducive to some social good – even if I am willing to accept the version of the good in question – is not the right kind of explanation to change my deep self intuitions, just as being told that somebody else has a 'better' aesthetic sensibility is unlikely to make me find different things beautiful.[112]

A second option would be to adopt a relativist approach to our deep self intuitions. Within a relativist framework, different deep self attributions would constitute propositions capable of being true or false, where the truth relevant conditions are specific to the attributer's context. A potentially useful analogy is our etiquette discourse: there *is* a fact of the matter, in any given context, about what is polite or rude, and so etiquette claims can be true or false relative to a particular context. This is the kind of position that philosophers like Finlay take with respect to moral claims in general: they reject the idea that our moral judgments include any presupposition that the moral values they involve have a kind of 'absolute' authority, and argue that even if they did this wouldn't "contaminate the meaning of truth conditions of moral claims" (Finlay, 2008).

How might we imagine the kind of background relative to which a given deep self attribution could be true? One plausible way of cashing this out would be to say that a given deep self attribution is true when it represents the most parsimonious explanation available within a given subject's underlying folk-psychological model. The truth conditions for deep self attributions would therefore be relative to each individual's folk-psychological model such that deep self attributions are almost always true (subject, perhaps, to revisions where some feature of a situation went initially unnoticed by the attributing agent). On this kind of account it makes less sense to talk about truth conditions than about "relational application conditions for moral concepts" (Finlay, 2008). Essentially, whilst there might not be a fact of the matter about a particular agent's deep self, there is a fact of the matter about what is an appropriate deployment of the deep self concept relative to the background of a particular folk-psychological model. This relativist approach highlights the functional nature of deep self

[112] The best that can be hoped for is perhaps that people succeed in 'suppressing' their deep self intuitions – refusing to express them in the kinds of blaming or praising responses that they ordinarily entail.

attributions within our folk-psychology over any potential assumptions being made about the referent of such claims.

If the relativist account reduces truth conditions to the correct (functional) application of moral concepts, it is tempting to imagine a way in which this might give us a foothold on conflicting deep self attributions. We might attempt, for example, to compare the folk-psychological models in which each deep self attribution is located in terms of their relative predictive accuracy. The problem is that how predictively accurate any particular model ends up being will always depend on the data with which it is presented. Whilst we might be able to come up with some kind of global standard of 'normal' human behaviour against which to compare the predictions of different folk-psychological models, members of any particular community would be right to reject such a standard. They would rightly refuse to accept that the 'better' deep self attributions are those that emerge from a folk-psychological model that more closely represents something like a global descriptive average of human behaviour. The reason is that, with respect to the subset of overall behaviour with which they are concerned in their everyday lives, the global, 'objective' model of human behaviour would be less predictively useful than their own, local models.

When faced with competing deep self attributions people are also unlikely to simply accept the relativist proposition that each is true in its own context. This is because we don't thing of our deep self propositions as including the kind of indexicals that relativists insert in them. The important difference between deep self talk and our etiquette discourse is that most subjects, when making etiquette claims, *understand* that the claim they are making has its truth conditions in some kind of local consensus about what is polite or rude. The average person, I would suggest, does not make deep self attributions in the same way. There is no kind of appeal to consensus – even local consensus – that is likely to change anyone's mind about their deep self attributions in the same way that it might in our etiquette discourse. The realist assumption, it seems, *is* a feature of our deep self talk.

The option that we are left with is to take an error-theory approach to our deep self attributions. This involves framing our deep self attributions as propositions – capable of being true or false – that are simply always *false*. It is the kind of position that philosophers like Joyce have taken to moral judgments in general, claiming that as propositions they do involve claims

about an objective categorial imperative and that such claims are always false.[113] I will argue that this is the most useful way of thinking about deep self attributions: when we make them we not only expect that other agents should see the 'truth' of the claims in question, but that they should do so independently of any particular moral institutions or normative frameworks. Based on the arguments that I have presented in Chapters II and III, I think it is most realistic to view this kind of realist claim in our deep self attributions as false.

What is interesting about the moral error theory approach is that it leaves us with an important practical choice, between a fictionalist and an eliminativist approach to our deep self attributions – essentially asking whether the fiction of the deep self is ultimately a useful one. The former approach involves characterising our deep self talk as a kind of useful fiction – one that brings about certain desirable consequences through its influence on our behaviour – which we ought to maintain despite the error it involves. The latter involves arguing that we would be best to eliminate the error in question and do away with (realist) deep self talk altogether. The fictionalist approach is likely to appeal to those who endorse a utilitarian view of our blaming practices. It is most consonant with the explanation of the cognitive mechanism behind our deep self attributions that involves some way of identifying those cases in which it would be 'useful' to praise or blame particular agents. The mechanism behind our deep self intuitions might be explained as having evolved specifically to be sensitive to indicators of the kind of situation in which blaming or praising agents was somehow fitness-enhancing for our ancestors.[114] Examples of this kind of fictionalist approach include the works of Doris (2015) and Alfano (2013), both of whom see the 'fiction' of moral character as an important way of promoting moral or prosocial behaviour.

I am inclined to think that the fictionalist account of deep self attributions is quite consistent with the explanation of the cognitive mechanism behind our deep self intuitions that

[113] See, for example, (Joyce, 2001, 2011). As he highlights, one important distinction between the relativist and the error-theorist about moral judgments is that whilst the relativist thinks of our everyday moral claims as involving *subjective* categorical imperatives (which may be made legitimate by the presence of particular moral institutions) the error-theorist sees such claims as involving *objective* categorical imperatives: claims about what constitutes a reason for action for every agent both independently of her desires or ends and without reference to any kind of moral institution.

[114] While we have only been concerned with attributability judgments – which do not themselves immediately translate into blaming responses or punishments – the importance of attributability judgments as a prerequisite for any blaming responses makes such a utilitarian explanation plausible. It is perhaps even plausible that the mechanism for attributing (fictional) deep selves operates even in situations in which there might be other practical reasons (or accountability-based reasons) for the agent *not* to blame the other, if indeed the mechanism in question has been constructed on the basis of a reliable but not exact correspondence between attributability judgments and blaming (or praising) responses.

I have defended. Within a given social context it seems likely that the mechanism I have described may well be a good way of identifying those norm-violations that it makes most sense to reward or punish in order to maintain normal standards of behaviour. And, as I have suggested before, it seems likely that against a particular background of 'normal' behaviour this is an effective way of coming up with causal explanations that form predictively accurate folk-psychological models. But how does it equip us to make sense of conflicting deep self attributions? Even if it were possible to say which deep self attributions are the most conducive to improving behaviour generally (which seems unlikely given the quite serious variability in conceptions of what 'better' behaviour would look like) I am inclined to agree with Strawson about the likely impact that this kind of explanation would have on people's actual deep self attributions:

> It is far from wrong to emphasise the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices. What is wrong is to forget that these practices, and their reception, the reactions to them, really are expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them. Indeed, the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this. (Strawson, 1962)

Essentially, according to Strawson, the reactive attitudes (or, on my account, our deep self intuitions) are products of our natures, such that it is not open to us to change them rationally, on account of the utilitarian merits of some alternative: "it is useless to ask whether it would not be rational for us to do what it is not in our nature to (be able to) do" (Strawson, 1962).[115] In this sense the fictionalist's suggestion that we ought to keep our deep self attributions because of some utilitarian benefit is quite unnecessary: we simply cannot change the cognitive mechanisms that select the causal explanations underpinning our deep self attributions. The kind of argument that the fictionalist deals in would at most be capable of convincing us to modify our *expressions* of the judgments in question, perhaps withholding our blaming or praising responses in certain cases. For this reason I would suggest that a fictionalist approach, with its focus on the 'utility' of deep self attributions, also doesn't provide us with a particularly useful way of looking at conflicting deep self attributions.

The remaining option for the error-theorist – the eliminativist approach – would be to bypass talk about the deep self altogether and instead deal directly with the underlying

[115] Strawson here was referring to the supposed 'rationality' of giving up our reactive attitudes in light of the truth of the thesis of determinism, but the point applies equally to the question of changing up our deep self attributions because some other set of attributions would be more rational in light of some utilitarian goal.

differences that produce competing deep self intuitions in the first place. Whilst there is obviously little chance of convincing the folk to abandon their deep self talk intuitions in their everyday interactions, I will argue that eliminating deep self talk from our theoretical discussion – not just as philosophers, but also as social scientists and policy-makers – will also provide us with the most useful frame for considering competing deep self intuitions. In order to show why this is the case, I now turn to a case-study of our deep self attributions around drug use and addictive behaviour.

§3) Drug Use – A Case Study

The unwilling and willing addicts are one of the central examples around which Frankfurt and many others after him have constructed theories of the deep self and moral responsibility.[116] The centrality of this particular example – of drug use, or addictive behaviour – appears not to have struck readers of the deep self literature nearly as much as I think it ought to have. The reason why it ought to strike us, certainly far more than it seems to, is that the apparently quite central 'model' of human behaviour that is the unwilling addict is, historically speaking, a relatively recent phenomenon. Put simply, the folk-imagination has not always contained the figure of an unwilling addict – not until the 'disease model' of addiction came along and portrayed the addict as the victim of a chemical process resulting in desires that he is powerless to resist. Now, by contrast, the idea of the unwilling addict is so widely accepted that more recent authors like Sripada (2017) have felt the need to explain that, though we might find the idea of a *willing* addict wildly implausible (even to the point of assuming that anybody who *claimed* to be a willing addict must be 'self-deceived or otherwise mistaken') it is, at the very least, a *metaphysical* possibility such that we may stipulate its existence for the purposes of a philosophical thought experiment. Surely, in historical perspective, this must strike us as a quite impressive reversal. What I aim to present in this section is a comparison of two historical models of addiction and an explanation of how a reversal like this can take place. I will then consider an emerging model of addiction and ask how we might assess it, as well as how, if it strikes us as an improvement on the alternatives, we might go about convincing people of it.

The underlying analysis that I will be relying on comes from Pickard (2017) who presents three different models of addiction: the 'moral model', the 'disease model' and the emerging 'new choice model'. The moral model, described by Pickard as the dominant model of addiction

[116] See (Frankfurt, 1971; Sripada, 2017b; Watson, 1975) as just a few examples.

in the first half of the twentieth century,[117] has two important features: (i) it views drug use as a choice, and (ii) it explains the choice made by drug users in terms of some kind of moral failing. The disease model, with which most people today are likely familiar, involves characterising the behaviour of addicts as compulsive, such that "addicts literally cannot help using the drugs, and have no choice over consumption" (Pickard, 2017). It is the disease model that provides us with our image of the 'unwilling addict'. The new choice model identified by Pickard involves characterising drug use as a choice, but one that doesn't require us to postulate any kind of moral defect in order to explain. Each involves a different causal explanation for drug use that is produced against the background of a different folk-psychological model, each influenced by a different set of social narratives about drug use.

The moral model of addiction, returning to our model-theoretic explanation of deep self intuitions, begins by approaching drug use as a kind of norm-violating behaviour: taking drugs is now how we would expect ordinary people to behave, and so we need to come up with some kind of difference-maker explanation for this unusual behaviour. What drug use involves, according to the relevant social narratives, is a purely hedonic pleasure, enjoyed to the detriment of all other activities. The moral model of addiction, in its context, is probably reasonably parsimonious as a causal explanation of this behaviour: it involves attributing to drug users selfish, anti-social inclinations of a kind that likely already constitute an existing 'type' within the underlying folk-psychological model.[118] The assumption that needs to be made on the moral model of addiction – that some group of people has (norm-violating) selfish, anti-social motivations – doesn't involve any serious deviation from the prior probability assigned to people having this kind of motivation generally. The consequence of this moral model of addiction is that addicts are to be blamed for their behaviour as it is reflective of their deep selves, and both the stigmatisation of addiction and a range of practices of punishment and ostracism flow from this.

How do we get from the moral model of addiction to the disease model that most of us are likely familiar with today? Is it simply driven by the emergence of empirical 'evidence' from studies on the neurobiology of addiction? This alone is not a sufficient explanation, as there is no necessary link between facts about neurological mechanisms and our deep self

[117] See (Pickard, 2017, p. 170). Whilst empirical research into such attitudes is comparatively recent, contemporary studies have identified two main competing folk models of addiction, sometimes described as the 'moral' and the 'medical' – see, for example, (Henderson & Dressler, 2017).
[118] Perhaps there is even a further story to be told here about how the mental representation of the addict 'type' conveniently adapted existing stereotypes of racial or class differences.

attributions, though of course it plays some role. Other important factors no doubt also include the changing face of the drug addict in the public imagination. As the stigma attached to drug use has increased, the social 'image' of the drug addict has come to reflect this: the 'junkie' of our social narrative is socially ostracised, dirty, suffering various health problems – in short, suffering. This provides us with a different background against which to explain the norm-violating behaviour in question: the idea that addicts use drugs out of some simple selfish desire for hedonic pleasure becomes less and less plausible, and the kind of underlying motivations we would need to posit in order to explain the behaviour in question as a choice become increasingly unparsimonious. It is highly improbable, according to our underlying folk-psychological model, that anyone would *choose* the obviously terrible life that drug use involves. And so a different kind of explanation – in this case a relatively new explanation in terms of the addictive properties of drugs – becomes more plausible. Because the explanation in question does not refer to any underlying (norm-violating) disposition in the agents in question, the disease model of addiction does not involve attributing any blame to drug addicts.[119]

The interesting possibility raised by Pickard is of a third model of addiction that characterises drug use as a choice but doesn't require an explanation of that choice in terms of moral defects – a model that doesn't involve attributing norm-violating deep motivations to the addicts in question. Pickard begins by providing us with the characterisation of addictive behaviour that provides the background for this model:

> To briefly review some of this evidence: Anecdotal and first-person reports abound of addicts (including those with a DSM-based diagnosis of dependence) going "cold turkey". Large scale epidemiological studies demonstrate that the majority of addicts "mature out" without clinical intervention in their late twenties and early thirties, as the responsibilities and opportunities of adulthood, such as parenthood and employment, increase. Rates of use are cost-sensitive: indeed some addicts choose to undergo withdrawal in order to decrease tolerance, thereby reducing the cost of future use. There is increasing evidence that Contingency Management treatment improves abstinence and treatment-compliance, compared to standard forms of treatment such as counselling and cognitive-behavioural therapy, by offering a reward structure of alternative goods, such as modest monetary influences and small prizes, on condition that addicts produce clean urine samples.

[119] A more common way in which the emergence of the disease model is understood is as "a call for compassion and a force for social justice and good" (Pickard, 2017) – which explains the transition in terms of a kind of 'motivated cognition': people *wanted* to stop blaming addicts and adopt a more compassionate approach and so adopted the beliefs (the deep self attributions) that were consistent with that. As I have suggested before though, this kind of desire for social change simply doesn't seem like the right kind of thing to make us change our deep self attributions. It also seems to get the order of explanation wrong: it is not by virtue of the socially desirable consequences that it may have that a particular model of behaviour becomes plausible, but rather the case that *in light of* the plausibility of a given model of behaviour certain social practices *become* desirable.

> […] Finally, since Bruce Alexander's seminal experiment "Rat Park" first intimated that something similar might be true of rats, animal research on addiction has convincingly demonstrated that, although the majority of cocaine-addicted rats will escalate self-administration, sometimes to the point of death, if no alternative goods are available, they will by contrast forego cocaine and choose alternative goods, such as saccharin or same-sex snuggling, if available. In short, the evidence is strong that drug use in addiction is not involuntary: addicts are responsive to incentives and so have choice and a degree of control over their consumption in a great many circumstances. (Pickard, 2017, pp. 171–172)

Essentially what she is suggesting is that patterns of drug use are what we would expect to see for agents making a choice. This makes the explanation in terms of the addictive properties of drugs slightly less parsimonious, as the much simpler explanation of behaviour that follows the pattern of other choices is that it too is a choice. Pickard is also suggesting that there is an *objective* model of addictive behaviour against which this model is in fact more predictively accurate. So how do we avoid the conclusion that this apparently norm-violating choice is best explained by norm-violating deep motivations in the agents in question? The answer is by changing the social narrative about the choice in question so that we can explain it as a normal response for agents in a particular kind of situation. Pickard starts by explaining that the effects of drug use are something that it is quite natural to desire:

> Throughout human history, drugs have been used to achieve a host of valuable ends, including at minimum the following: (1) improved social interaction; (2) facilitated mating and sex; (3) heightened cognitive performance; (4) facilitated recovery and coping with stress; (5) self-medication for negative emotions, psychological distress and other mental health problems and symptoms; (6) sensory curiosity – expanded experiential horizon; and, finally, (7) euphoria and hedonia – in other words, pleasure. Drugs make us feel good, provide relief from suffering, and help us do various things we want to do better. (Pickard, 2017, p. 177)

At the same time she explains how the choice of these over other desirable goods may be perfectly rational in a certain context of options:

> [Drug users] may experience extreme psychological distress alongside a host of mental health problems apart from their addiction, feel a lack of psychosocial integration, and are at a socio-economic disadvantage such that they have severely limited opportunities. These circumstances are central to understanding addiction in many contexts. Put crudely, the reason is simply that drugs offer a way of coping with stress, pain, and some of the worst of life's miseries, when there is little possibility for genuine hope of improvement and limited alternative goods on offer. In such circumstances, whatever harms accrue from using drugs must be weighed against whatever harms accrue from not using them. For this reason, the explanation of addiction and its associated negative consequences must lie in no small part with the psycho-socio-economic circumstances that cause such suffering and limit opportunities. (Pickard, 2017, p. 178)

Against this background understanding of the situation that many drug users are faced with, the behaviour in question no longer represents a deviation from what our underlying folk-psychological model would predict an agent might do in that situation.

§4) Conclusion

How are we to go about assessing the different deep self attributions that these three models represent? The expressivist and the relativist approaches, as mentioned above, fail to capture the sense in which the different deep self attributions in question are actually in conflict with one another: members of the same society disagree about which model represents the *truth* about drug users' deep selves, and it is not a disagreement which we (or they) can simply ignore because important questions of social policy depend on its resolution. There are ways of framing, within a relativist or expressivist account, why one deep self attribution might be more desirable than another – expressed, for example, in terms of the kind of deep self sensibility that is more conducive to some agreed-upon good, or of an 'objective' background against which one model is more predictively accurate. But these arguments are unlikely to change people's deep self intuitions.

The fictionalist – who argues that deep self talk in general is valuable in some consequentialist sense – may also have an explanation as to which particular deep self attributions are optimal from this utilitarian perspective, but I doubt that many people would be capable of revising their deep self intuitions (if my model of the cognitive mechanisms underlying those intuitions is accurate) on the basis of this utilitarian argument. Even more problematic is the fact that people's competing deep self intuitions are likely to make it difficult to agree on any model of what is optimal in a utilitarian sense to begin with. One fictionalist position with respect to the set of contested deep self intuitions above might be that we ought to conform our deep self attributions to the new choice model because that model leads to more successful rehabilitation of addicts.[120] Someone who endorses a moral model of addiction (who attributes drug use to a moral failing) might see rehabilitation as a good to be promoted, but also see just blame and just punishment (based on the deep self attributions involved in the moral model) as similarly important goods. For this reason the fictionalist is likely to face difficulty in finding common ground as the basis for a fictionalist argument about which set of

[120] Pickard's claim, for example, is that the new choice model of addiction is not only the most clinically effective way of achieving rehabilitation but also the most likely to lead to effective social policy to reduce drug-associated harms.

deep self attributions is most useful – people who make different deep self attributions will not have the same understanding of the 'ends' that our approach to addition ought to achieve. This kind of thought – about the inherent *value* of treating agents according to their moral desert – is what would make it quite difficult for most people to accept a utilitarian justification for adopting what they see as a fiction about others' deep selves.

The eliminativist approach, that I consider preferable, is as follows. Our deep self talk contains within it a realist assumption that is simply false. There are no psychological facts about any agent capable of justifying any attributability judgments in the way that our ordinary deep self talk implies, and so we ought to avoid such deep self talk altogether. At the same time, deep self intuitions are the inevitable product of the underlying cognitive mechanisms of our folk psychology – a part of our nature that we cannot change by any voluntary, conscious act addressed at the intuitions directly. Our deep self talk is flawed, but as folk-psychologists we can't help viewing the behaviour of those around us in terms of deep selves. This obviously implies that it is impossible to eliminate deep self talk altogether – but is also suggests that it is useless to try and convince people, on the basis of some abstract justification, to modify their deep self attributions. Arguments of this kind – of the kind that I have suggested the fictionalist or the relativist might try and provide, that address deep self claims directly and attempt to provide *reasons* for changing them – are unlikely to change anyone's mind. The eliminativist position that I would advocate for the social activist is to eliminate the deep self concept from *their own* discourse, rather than providing the folk with theoretical justifications for doing so. By avoiding deep self talk altogether the moral reformer avoids running up against the realist assumption that accompanies people's deep self attributions. By focusing instead on developing the kind of social narratives that underpin particular deep self attributions, it might actually be possible to change them.

§5) Bibliography

Adams, R. M. (1985). Involuntary Sins. *The Philosophical Review*, 94(1), 3.

Alfano, M. (2013). *Character as moral fiction*. Cambridge University Press.

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in Knobe-effect cases: A unified account of the data. *The Monist*, 95(2), 264–289.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63, 368.

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science*, 10(6), 790–812.

Andow, J. (2015). Expecting Moral Philosophers to be Reliable. *Dialectica*, 69(2), 205–220.

Arpaly, N., & Schroeder, T. (2013). *In Praise of Desire*. OUP USA.

Bear, A., & Knobe, J. (2016). Normality: Part descriptive, part prescriptive. *Cognition*.

Beebe, J. R. (2013). A Knobe Effect for Belief Ascriptions. *Review of Philosophy and Psychology*, 4(2), 235–258.

Beebe, J. R. (2016). Do bad people know more? Interactions between attributions of knowledge and blame. *Synthese*, 193(8), 2633–2657.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, 25(4), 474–498.

Beever, A. (2013). The Declaratory Theory of Law. *Oxford Journal of Legal Studies*, 33(3), 421–444.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, 33(3), 169–185.

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy*, 49(188), 123–134.

Bennett, J. (1980). Accountability. In Z. van Straaten (Ed.), *Philosophical Subjects: Essays Presented to P.F. Strawson* (pp. 14–47). Oxford: Clarendon.

Broadus, A. D., & Evans, W. P. (2015). Developing the public attitudes about addiction instrument. *Addiction Research & Theory*, 23(2), 115–130.

Brownstein, M. (2016). Attributionism and Moral Responsibility for Implicit Bias. *Review of Philosophy and Psychology*, 7(4), 765–786.

Buckwalter, W., & Stich, S. (2014). Gender and Philosophical Intuition*. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy: Volume 2*.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the Belief in Good True Selves. *Trends in Cognitive Sciences*, 21(9), 634–636.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2017). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science* 42(1), 134-160.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2016). Normative Judgments and Individual Essence. *Cognitive Science*, 41, 382–402.

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, 29(1), 1–22.

Doris, J. M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Downey, L., Rosengren, D. B., & Donovan, D. M. (2000). To thine own self be true: Self-concept and motivation for abstinence among substance abusers. *Addictive Behaviors*, 25(5), 743–757.

Finlay, S. (2008). The error in the error theory. *Australasian Journal of Philosophy*, 86(3), 347–369.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, 66(23), 829–839.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1), 5.

Frankfurt, H. G. (1982). The Importance of What We Care About. *Synthese*, 53(2), 257–272.

Frankfurt, H. G. (1987). Identification and Wholeheartedness. In F. D. Schoeman (Ed.), *Responsiblity, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press.

Frankfurt, H. G. (1988). Rationality and the Unthinkable. In *The importance of what we care about: Philosophical essays*. Cambridge University Press

Frankfurt, H. G. (1992). The Faintest Passion. *Proceedings and Addresses of the American Philosophical Association*, 66(3), 5–16.

Frankfurt, H. G. (1998). On the Necessity of Ideals. In *Necessity, Volition, and Love* (pp. 108–116). Cambridge, U.K. ; New York: Cambridge University Press.

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford, New York: Oxford University Press.

Gelman, S. A., & Medin, D. L. (1993). What's so essential about essentialism? A different perspective on the interaction of perception, language, and conceptual knowledge. *Cognitive Development*, 8(2), 157–167.

George Sher. (2005). *In Praise of Blame*. Oxford: Oxford University Press.

Greene, J. D. (2008). The secret joke of Kant's soul. *Moral Psychology*, 3, 35–79.

Greenwald, G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480.

Haidt, J. (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgement. *Psychological Review*, 108(4), 814–834.

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, 26(2), 201–218.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, 30(12), 1661–1673.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity. *Cognitive Science*, 41(3), 744–767.

Henderson, N. L., & Dressler, W. W. (2017). Medical Disease or Moral Defect? Stigma Attribution and Cultural Models of Addiction Causality in a University Population. *Culture, Medicine, and Psychiatry*, 41(4), 480–498.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, 106(11), 587–612.

Hobbes, T. (1997). *Leviathan* (R. E. Flathman & D. Johnston, Eds.). New York: W.W. Norton.

Hume, D. (1978). *An Enquiry Concerning Human Understanding* (P. H. Nidditch, Ed.). Oxford: Clarendon Press.

Joyce, R. (2001). *The Myth of Morality*. Cambridge University Press.

Joyce, R. (2011). The Error In 'The Error In The Error Theory.' *Australasian Journal of Philosophy*, 89(3), 519–534.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28(2), 107.

Knobe, J. (2003a). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63(3), 190–194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16(2), 309–324.

Knobe, J. (2007). Reason Explanation in Folk Psychology. *Midwest Studies In Philosophy*, 31(1), 90–106.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33(04), 315–329.

Knobe, J., & Burra, A. (2006). The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study. *Journal of Cognition and Culture*, 6(1), 113–132.

Knobe, J., & Fraser, B. (2008). Cause and Moral Judgment: Two Experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, MIT Press.

Knobe, J., & Mendlow, G. S. (2004). The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology. *Journal of Theoretical and Philosophical Psychology*, 24(2), 252.

Knobe, J., & Nichols, S. (2017). Free Will and the Bounds of the Self. In R. Kane (Ed.), *Oxford Handbook of Free Will*. New York: Oxford University Press.

Knobe, J., & Preston-Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues*, 19(1), 131–147.

Leben, D. (2014). When psychology undermines beliefs. *Philosophical Psychology*, 27(3), 328–350.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect: Theory of Mind and Moral Judgment. *Psychological Science*, 17(5), 421–427.

Liao, S. (2016). Are philosophers good intuition predictors? *Philosophical Psychology*, 29(7), 1004–1014.

Machery, E. (2004). Semantics, cross-cultural style. *Cognition*, 92(3), B1–B12.

MacIntyre, A. (1984). *After virtue: A study in moral theory / by Alasdair MacIntyre*. Notre Dame, Ind: University of Notre Dame Press.

Malle, B. F. (2005). Self-Other Asymmetries in Behavior Explanations. *Self IN Social Judgment*, 155.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895–919.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, 33(2), 101–121.

McKenna, M., & Coates, D. J. (2018). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Medin, D. L., & Ortony, A. (1989). Psychological Essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge ; New York: Cambridge University Press.

Menzies, P., & Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, 44(2), 187–203.

Meyers, D. T. (1987). Personal autonomy and the paradox of feminine socialization. *The Journal of Philosophy*, 84(11), 619–628.

Mizrahi, M. (2015). Three arguments against the expertise defense. *Metaphilosophy*, 46(1), 52–64.

Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93, 1–17.

Murray, D., & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, 88(2), 434–467.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9(2), 203–219.

Nadelhoffer, T., & Feltz, A. (2008). The Actor–Observer Bias and Moral Intuitions: Adding Fuel to Sinnott-Armstrong's Fire. *Neuroethics*, 1(2), 133–144.

Nelkin, D. K. (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, 40(2), 203–216.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs About the True Self Explain Asymmetries Based on Moral Judgment. *Cognitive Science*, 39(1), 96–125.

Nichols, S. (2017). The Rationality of Psychological Essentialism. In *Advances in Experimental Philosophy*. *Experimental Metaphysics* (pp. 117–134). Bloomsbury.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, 41(4), 663–685.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17(3), 145–171.

Phillips, J., & Knobe, J. (2009). Moral Judgments and Intuitions About Freedom. *Psychological Inquiry*, 20(1), 30–36. h

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The Ordinary Concept of Happiness (and Others Like It). *Emotion Review*, 3(3), 320–322.

Pickard, H. (2017). Responsibility without Blame for Addiction. *Neuroethics*, 10(1), 169–180.

Ryberg, J. (2013). Moral intuitions and the expertise defence. *Analysis*, 73(1), 3–9.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2014). Folk Moral Relativism*. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy: Volume 2*. Oxford University Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition*, 20(4), 1722–1731.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27(2), 135–153.

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, 141, 127–137.

Shoemaker, D. (2003). Caring, Identification, and Agency. *Ethics*, 114(1), 88–118.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121(3), 602–632.

Shoemaker, D. (2015a). Ecumenical Attributability. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility: New Essays*. Oxford University Press.

Shoemaker, D. (2015b). *Responsibility From the Margins*. Oxford University Press.

Shoemaker, D. (2017). Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to BlameResponse-Dependent ResponsibilityDavid Shoemaker. *The Philosophical Review*, 126(4), 481–527.

Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2). MIT Press.

Smith, A. M. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*, 115(2), 236–271.

Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics*, 11(4), 465–484.

Smith, A. M. (2008). Control, Responsibility, and Moral Assessment. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 138(3), 367–392.

Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics*, 122(3), 575–589.

Sripada, C. (2010). The Deep Self Model and Asymmetries in Folk Judgements About Intentional Action. *Philosophical Studies*, 151(2), 159–167.

Sripada, C. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, 48(1), 232–238.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, 173(5), 1203–1232.

Sripada, C. (2017a). At the center of agency, the deep self.

Sripada, C. (2017b). Frankfurt's Unwilling and Willing Addicts. *Mind*, 126(503), 781–815. h

Sripada, C., & Konrath, S. (2011). Telling More Than We Can Know About Intentional Action. *Mind & Language*, 26(3), 353–380.

Stich, S. P., & Nichols, S. (2003). Folk Psychology. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell Guide to Philosophy of Mind* (pp. 235–255). Blackwell Publishing.

Stich, S., & Weinberg, J. M. (2001). [Review of *Jackson's Empirical Assumptions*, by F. Jackson]. *Philosophy and Phenomenological Research*, 62(3), 637.

Strawson, P. F. (1980). Replies. In *Philosophical Subjects: Essays Presented to P.F. Strawson* (pp. 260–296). Oxford: Clarendon Press.

Strawson, Peter F. (1962). Freedom and Resentment. In G. Watson (Ed.), *Proceedings of the British Academy, Volume 48: 1962* (pp. 1–25). Oxford University press.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159–171.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, 26(9), 1469–1479.

Taylor, C. (1992). *The Ethics of Authenticity*. Harvard University Press.

Tobia, K. P. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75(3), 396–405.

Tobia, K. P. (2016). Personal Identity, Direction of Change, and Neuroethics. *Neuroethics*, 9(1), 37–43.

Tobia, K. P., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology*, 26(5), 629–638.

Todd, P. (2016). Strawson, Moral Responsibility, and the "Order of Explanation": An Intervention. *Ethics*, 127(1), 208–240.

Tognazzini, N., A. (2013). Blameworthiness and the Affective Account of Blame. *Philosophia*, 41, 1299–1312.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481), 453–458.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1), 87–100.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

Varga, S. (2011). Self-Realization and Owing to Others: An Indirect Constraint? *International Journal of Philosophical Studies*, 19(1), 75–86.

Wallace, R. J. (1994). Responsibility. In *Responsibility and the Moral Sentiments* (pp. 51–83). Harvard University Press.

Watson, G. (1975). Free Agency. *The Journal of Philosophy*, 72(8), 205–220.

Watson, G. (1977). Skepticism about Weakness of Will. *The Philosophical Review*, 86(3),

Watson, G. (1987a). Free Action and Free Will. *Mind*, 96(382), 145–172.

Watson, G. (1987b). Responsibility and the Limits of Evil. In F. D. Schoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology*. Cambridge University Press.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, 24(2), 227–248.

Watson, G. (2004a). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In *Agency and Answerability*. Oxford University Press.

Watson, G. (2004b). Volitional Necessities. In *Agency and Answerability*. Oxford University Press

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, 23(3), 331–355.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29(1/2), 429–460.

Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy*, 42(3), 215–229.

Wolf, S. (1981). The importance of free will. *Mind*, 90(359), 386–405.

Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46–62). Cambridge University Press.

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.