**Title**
On the Fringes of Moral Responsibility: Skepticism, self-deception, delusion, and addiction



**Author**
Gibson, Quinn Hiroshi

**Publication Date**
2017

Peer reviewed|Thesis/dissertation

**On the Fringes of Moral Responsibility: Skepticism, Self-deception, Delusion, and Addiction**

by

Quinn Hiroshi Gibson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Campbell, Co-chair
Professor R. Jay Wallace, Co-chair
Professor Tania Lombrozo

Summer 2017

**On the Fringes of Moral Responsibility: Skepticism, Self-deception, Delusion, and Addiction**

## Abstract

On the Fringes of Moral Responsibility: Skepticism, Self-deception, Delusion, and Addiction

by

Quinn Hiroshi Gibson

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor John Campbell, Co-chair

Professor R. Jay Wallace, Co-chair

This dissertation is a collection of essays under the theme of moral responsibility 'at the margins'. I investigate a number of examples of disordered agency and cognition — self-deception, delusion, and addiction — through the lens of a so-called 'reasons-responsiveness' theory of morally responsible agency, employing the theory to examine the extent to which agents in those conditions are morally responsible and in virtue of what this is so.

In Chapter 2, after a brief introductory chapter, and before getting into the individual disordered phenomena, I develop and defend the reasons-responsiveness theory of responsible agency to which I will appeal in later chapters. Such theories — according to which responsible agency is based in an agent's capacity for recognizing and responding to reasons for action — are not entirely new. However, developed in the right way, they are also well-equipped to respond to a kind of skeptical challenge to morally responsible agency that has somewhat recently come into vogue. This skeptical challenge is motivated by recent findings in social and cognitive psychology that seem to show that much of human behaviour is motivated by considerations which are, from the perspective of justifying action, irrelevant. For example, contributions to a communal office coffee fund can as much as triple when the instructions are accompanied with a pair of watchful 'eyes' on the wall. I argue that of all mainstream theories of agency, the reasons-responsiveness theory is least threatened by results such as these. I further respond by addressing a dispute between reasons-responsiveness theorists themselves: what is required for someone to count as responding to reason? I argue for a liberal interpretation of this requirement on independent grounds, and note that such a version of the theory is even better equipped to respond to the skeptic, yielding a theory of agency which is actually enhanced by appeal to the empirical results.

In Chapters 3 and 4 I develop a novel account of self-deception and use that account to address the question: Are some delusional subjects responsible for their delusions? The central difficulty for the philosophical theory of self-deception has been to yield a psychologically plausible description of its dynamics. But self-deception is also paradigmatically

intentional behaviour for which agents are typically blameworthy. I argue that no extant account of self-deception can capture both of these features. On my account, what makes a state a self-deceptive one is not determined by how it comes about. Rather, it is determined by how that belief is maintained. Self-deception, on this view, is a willful failure, a refusal, to meet epistemic requirements for motivationally biased reasons. Thus, self-deceivers are typically responsible for their self-deception. I further argue that if this account is correct, there will be at least some cases of delusion (e.g., the Reverse Othello and Capgras delusions) for which agents are, in some sense, responsible. Appealing to the distinction between blameworthiness and (what I shall call) 'attributability', I claim that this leads us not to the conclusion that delusional subjects should be blamed, but instead to a more nuanced understanding of the kind of agency involved in the dynamics of delusion, and of the reasons these subjects are excused.

The final chapter is about addiction. Perhaps the central question raised by addiction is: to what extent are addicts responsible agents? Theorists notoriously oscillate between two extreme positions: (1) that addicts are just like unimpaired agents and are fully responsible and (2) that addicts helplessly suffer a condition that leaves them utterly without self-control. I argue against both extreme positions, engaging with current science at both turns. Against (2), I argue that there is no satisfactory understanding of the 'brain disease theory' of addiction that entails that addicts are not responsible agents. I then argue against (1) by considering addicts at different stages of addiction — those who are aware of their predicament vs. those who are not (although they should be). With respect to the unaware, I argue that they share some features with the self-deceived which explains their insensitivity to a rationally circumscribed body of evidence. Concerning the aware, I appeal to empirical work on 'ego-depletion' and willpower — and to Chapter 2's theory of responsibility — to argue that these addicts suffer a graded impairment of the will, one that partially excuses them from blameworthiness.

For my parents, who didn't insist that I go to law school.

# Contents

# Acknowledgments

It is difficult to exhaustively acknowledge everyone to whom I owe a debt for making the writing of this dissertation possible. My interest in philosophy as both a branch of intellectual inquiry and a professional discipline was stoked early during my time as an undergraduate at the University of Calgary. I enrolled in philosophy classes my first semester there with some vague notion that it was the right thing for me. The more-or-less determinate notion that I had in mind only approximately corresponded to reality, but luckily what I did encounter was inspiring, and resonated enough with what I previously had in mind to stoke a sustained interest, nay, a passionate devotion, to what I regarded as an admirably clear-headed way to address oneself to some of the most interesting questions that there are. Among the many people during those years who were responsible for nurturing my nascent identity as a philosopher I must thank John Baker, Jeremy Fantl, and Richard Zach for supporting my graduate school applications and providing very valuable mentorship, and Allen Habib for excellent supervision of my undergraduate honours thesis, and ongoing support since. I also thank David Dick and Ish Haji for, in their very different ways, assuring me I was doing the right thing.

Since I arrived at UC Berkeley in 2010 I have been blessed with the opportunity to spend seven or so years in the company of many excellent people, and it would impossible to name everyone in the Berkeley philosophy community to whom I owe a debt for lending an ear, being a sounding board, or providing innumerable other forms of often intangible support. Though I am sure to forget some, this group includes: Katherine Ammirati, Austin Andrews, Brian Berkey, Joseph Bjelde, Lara Buchak, Jeremiah Carey, Lisa Clarke, Klaus Corcilius, Sophie Dandelet, Caitlin Dolan, Amin Ebrahimi, Peter Epstein, Jessica Gelber, Kelly Glover, Nick Gooding, Tyler Haddow, Jim Hutchinson, Zac Irving, Julian Jonker, Dan Khokhar, Niko Kolodny, Richard Lawrence, Dylan Murray, Sven Neth, Antonia Peacocke, Emily Perry, Kirsten Pickering, Jens Pier, Rachel Rudolph, Umrao Sethi, Barry Stroud, Dave Suarez, Manuel Vargas, Justin Vlasits, Daniel Warren, and Yuan Wu.

Alex Kerr and Adam Bradley both deserve special mention for reminding me of the importance of the empirical for many of the questions that we philosophers hold dear. Ethan Jerzak equally deserves mention for never letting me forget the importance of conviction in a priori reasoning for the very same questions. Clara Lingle is to be thanked for never letting me become a fanatical devotee of either camp. All four are to be thanked for countless conversations, at every level of resolution, about the material in this dissertation. I also want to thank Clara for her love and encouragement, and for treating me with gentle good humour when my frustration with this project was showing.

Special thanks are also deserved by Mike Martin and Véronique Munoz-Dardé both of whom provided mentorship, support, and feedback on my work despite being outside of my committee.

I am also grateful to audiences at the 2015, 2016, and 2017 annual meetings of the Society for Philosophy and Psychology (at each of which were present strong contingents from the Gopnik and Lombrozo labs at Berkeley), the 2016 Eastern APA, and the 15th and 16th

annual Berkeley-London graduate conferences, for providing valuable feedback on much of the material in these pages.

Both of my advisors, Jay Wallace and John Campbell, provided excellent and consistent feedback and support throughout the entire dissertation writing process. We began meeting before I advanced to candidacy and in preparation for my qualifying exams, and since then I have met regularly with both of them (including by Skype if necessary). Over the years they have alerted me to countless errors and omissions, vastly improving the quality of the arguments contained herein, and my philosophical thinking generally.

# Chapter 1

# Introduction

This dissertation is a collection of essays about moral responsibility. Unlike traditional philosophical discussions about moral responsibility, however, these essays do not focus on the compatibility or incompatibility of freedom and determinism. Neither the notion of freedom nor the notion of determinism plays a large role in any of the discussions contained herein. Some of the reasons for this are negative. I do not find much use for any notion of freedom or of an action freely undertaken (as that notion sometimes figures in the idea of 'free will', for example) which builds in anything more than the idea that certain kinds of responses are appropriate in response to that action. In addition, I am of the Strawsonian (Strawson 1962) persuasion in thinking that the general practice of responding to each other with moral emotions and other reactions such as praise and blame is utterly indispensable to any life which is recognizably human. The task that remains, on this way of thinking about it, is to get as clear as we possibly can about which thoughts and actions justifiably occasion which sorts of responses. Thinking that free will is just that property possessed by an agent in virtue of which *any* such responses to her are ever appropriate, but that *some* such responses will be required to make for a recognizably human form of life robs the notion of free will of much of its independent interest. It diminishes the urgency of providing a justification, given from outside our responsibility practices, for those very practices themselves. In this sense, many of the questions raised in this dissertation (with an exception for some of those raised in Chapter 2) are raised internally to those practices.

More positively, however, I want to place the emphasis on the inside of our responsibility practices because I think there we can find many interesting and important questions about moral responsibility that either are often neglected or are only just beginning to come into focus as our empirically-based understanding of the mind — and its frequent less-than-ideal function — deepens. The positive and the negative considerations are importantly related. In my view, a philosophical concern with moral responsibility is, at bottom, a concern that our thoughts and feelings towards and about the people with whom we inhabit a shared social world be *appropriate*. And there are a great many cases where, because they are complicated, unfamiliar, or otherwise obscure, it is just not obvious how we ought to think about those involved. I take addiction and delusion to be examples of this phenomenon.

Anyone who has thought for more than a moment about either will recognize that there is more going on there than could possibly issue in a clear and obvious verdict about the extent to which the agents involved are morally responsible for their conduct. Here the task of moral philosophy is simply to dig in, try to reckon as much as possible with the empirical facts of the situation — what exactly is going on in the heads of these people; what their experiences are like; what they believe and what they desire — and bring to bear the best philosophical theories that we have to try to sift through the complexity. This is one of the things I have tried to do in the pages that follow.

Still, there are other cases where we are reasonably sure that our moral feelings (of disapproval, say) are appropriate, but it is not clear in virtue of what this is so. I take self-deception to be an example of this phenomenon. The concept of self-deception is not a morally neutral one. Built into the very idea of self-deception is the idea that it is somehow vicious conduct, that whatever it involves precisely, it is not something that an ideal agent would engage in. For all that, there is significant philosophical dispute over how best to understand its psychological dynamics. Some philosophers have gone so far as to say that there is no real psychological phenomenon which bears a sufficiently robust resemblance to our ordinary concept of self-deception to count as real self-deception at all. I take seriously that we should do our best to find a philosophical account that makes enough meaningful contact with the ordinary conception, but I also think that this requires that the account make sense of the normative aspects of the phenomenon, viz., that we think self-deception is vicious, or as I will say, blameworthy. The task here is different from the cases above. Here, we must try to come to a philosophical understanding of a complex mental phenomenon which preserves as much as possible the moral intuitions we bring pre-theoretically to bear on it. Here too, obviously, it is of the highest relevance what is going on 'in the heads' of subjects — what their experiences are like; what they believe and desire — and once again empirical facts are going to bear directly on how best to understand that.

Reckoning with the empirical is not a new phenomenon in philosophy, nor even in moral philosophy. But it has seen some renewed interest – and as I said, in a form different from the traditional questions about determinism — in recent years. The discussion in Chapter 2 is closest to this main stream of interest. I use Chapter 2 to introduce a few key notions that will run through the subsequent chapters. The first of these is so-called 'dual-process theory'. Very broadly speaking, dual-process theory is a theory, or family of theories, in empirical psychology about high-level cognitive architecture. According to this theory (at least on one way of understanding it) there are two fundamentally different types of cognitive processes. One type includes the conscious episodes of deliberative reasoning that we are all familiar with and which we likely bring to mind when we think about what thinking, paradigmatically, is. However, much of the cognitive activity that we are engaged in on a daily basis is not of this type. Rather it happens automatically, below the level of awareness, and it is systematically biased in a number of important ways. Moreover, its effects don't just crop up under experimental conditions. Much high-level thought and behaviour is influenced by it.

These unconscious processes, called 'type-1' processes, have been cropping up in empirical

psychology in one form or another for more than fifty years. But only fairly recently has it become clear that they together constitute a whole subterranean cognitive system. I begin Chapter 2 by reviewing some of the key findings that have led to this conclusion. Of course, not all psychologists (and certainly not all philosophers) are in agreement about the significance of the experimental findings. However, I am impressed by the convergence and momentum exhibited by this research program, and this is something that I make an effort to bring out.

The overall goal of Chapter 2 is to motivate, against the backdrop of these empirical results, the theory of morally responsible agency that I will appeal to, or which will loom large in the background, throughout the remainder of the dissertation. I have chosen to introduce the theory in this way for a couple of reasons. First, some philosophers and cognitive scientists have thought that the findings from empirical psychology make trouble for some traditional philosophical theories of morally responsible agency. I do not deny that there are some theories that might be imperiled by the results, but it is important that we formulate any such skeptical challenge carefully so as to be as clear as we possibly can be about precisely which theories the empirical results are supposed to make trouble for. In the end, I argue that the theory which I prefer is not imperiled by the empirical results. But I also argue that, suitably understood, some of the theories which have been thought to be imperiled are not imperiled either. This allows me to address more generally the significance of empirical results for philosophical theorizing about morally responsible agency at the same time as I introduce the theory that I prefer.

The second reason I haven chosen to introduce my preferred theory in this way is that I think it is nicely highlights the advantages of that theory, and it does so in a dialectically convenient and powerful way. The traditional philosophical theories that have become targets of the skeptical challenge have been grouped together under a particular heading, 'reflectivism'. I argue that someone who endorses a reflectivist theory has the resources to respond to the skeptical challenge (suitably understood), but I then argue that my preferred theory is even better-positioned to respond to the same challenge *and* that it better captures the phenomenon of morally responsible agency. My response to the skeptical challenge thus should be of interest even to someone who is not inclined towards my preferred theory, but further, my argument for that theory should be of interest to someone who is not particularly interested in empirically-based skepticism, but merely in how two different families of theories of morally responsible agency stack up against one another.

The theory that I prefer is a so-called 'reasons-responsiveness' theory of morally responsible agency. According to a theory of this kind, what makes someone a morally responsible agent is her capacity to recognize and respond to reasons for action. These theories are not entirely new, and have been extensively developed by, among others, Wallace (1994) and Fischer and Ravizza (1998). Central to many versions of the reasons-responsiveness theory is the idea of volitional self-control in response to reflective judgement. That is, many reasons-responsiveness theorists have emphasized the human capacity for bringing conduct into conformity with what we judge we ought to do. I don't mean to downplay the importance of this capacity. However, I also want to argue (with Arpaly 2001) that the notion of

volitional self-control can be extended to cases where the agent doesn't engage in reflective judgement. That is, there are ways of responding to reasons that do not require one to acknowledge those reasons consciously or explicitly. This will obviously help us in response to a skeptical challenge based in the idea that reflective judgement is doing less work than we thought it was. But it is also a very plausible extension of the theory: reasons, whatever they are precisely (and I prefer to think of them as facts), must impinge upon us somehow if the very idea of reasons-responsiveness is to make any sense at all. Paradigmatic examples of this seem to involve consciously recognizing reasons of a certain kind — reasons concerning what is rationally related to what, for example — in explicit judgement. But it should scarcely come as a surprise — especially against the empirically informed conception of cognition that I outline — that the facts which bear on what we ought to do can impinge upon us in other ways as well, some of which we may be scarcely aware of. It is tempting to think that when we act 'on our gut' or go with intuition that we are being foolhardy and ought to step back and take a moment in cool deliberation. This may be true — many decisions are regrettably taken too hastily. But, on the other hand, it may not be true. It depends how well tutored our emotions and our intuition are. There is a matter of fact about whether they are or are not attuned to the facts that bear on the deliberative situation. And if they are guiding us well, then being guided by them should count as being responsive to reasons. It may not be possible to tell, from the inside, whether our instincts should be trusted, but that is nothing but a familiar and often tragic fact about our epistemic limitations.

This feature of our deliberative lives thus has two faces. Agents who are set up to effectively respond unreflectively to normative reasons display a certain kind of highly admirable effortless virtue. This probably does not simply happen by accident, but rather by careful habituation. Aristotle was perhaps right when he observed that 'none of the virtues of character arises in us naturally' (*NE* II, 1, 19–20). And for those of us in whom such virtues have not (by habituation or otherwise) been inculcated, acting without rational intervention has its hazards. The hazards of so acting are one of the major themes of this dissertation. But the hazards only arise because sometimes, when we are not properly set up to respond to reasons unreflectively, we simply *must* exercise rationality if we are to act well. The suitably extended reasons-responsiveness theory that I will be working with identifies the capacities which are important for morally responsible agency with the whole suite of ways in which we can be, depending on our constitutions and the situations we find ourselves in, responsive to reasons. And it does so without privileging any of these capacities over the others.

Against this background understanding of the requirements for morally responsible agency I go on, in Chapters 3, 4, and 5, to investigate self-deception, delusion, and addiction, respectively. With self-deception, as I said above, the philosophical task is not so much to determine whether, in general, self-deceivers are responsible. It is rather to vindicate the judgement that they are so. I argue that the way to do this is to understand it as a culpable failure to do what is necessary to ensure that one is appropriately responsive to certain reasons for belief. That is, in self-deception, one is in an important sense unresponsive to reasons, but one *ought* to be (and can be) responsive to them. But saying this much is not enough. We must also give an account of self-deception that demystifies its psychological dy-

namics. Much philosophical discussion of self-deception has focused on this task, and it is an important one. What *is* self-deception? On the face of it, self-deception is puzzling: how can someone act so as to cause herself to intentionally believe something she already believes to be false? Any account of self-deception has to reckon with this. But the moral-psychological task of vindicating our responsibility judgements has received comparatively little attention. I take it that both of these are important desiderata for any theory of self-deception. I begin Chapter 3 by briefly taxonomizing some of the more well-known kinds of irrationality and by situating self-deception within that space. One of the most basic distinctions between forms of irrationality is between that which is motivated and that which is unmotivated. Typically, unmotivated irrationality can be explained without need to advert to any of the subject's desiderative states. The systematic biases that I discuss in Chapter 2 and that have, in large part, motivated the dual-processing research program, are responsible for irrationality which is, for the most part, of this kind. Motivated irrationality, on the other hand, is harder to cash out in subpersonal terms because it essentially involves some desire-like states — person-level states — of the subject. Wishful thinking and weakness of will are examples of irrationality of this kind (and I will draw parallels and connections with these two phenomena repeatedly throughout).

Self-deception is pretty clearly a form of motivated irrationality. Nevertheless, my account of self-deception, which I call 'Self-deception as Omisson', is inspired by the dual-process theory that has so effectively been wheeled in to explain so many of our cold biases. It is by appeal to this dual nature of the phenomenon that my view is able to both render a plausible account of the psychological dynamics of self-deception and to make sense of how self-deceivers are responsible for it. Many philosophers have been puzzled by how a subject can come to acquire a self-deceptive belief. It can seem as though, in order for the belief to count as properly self-deceptive, it would have to have been brought about intentionally. Not only does this lead to sticky issues about the connection between belief and the will, it seems especially problematic in the context where the subject is already thought to knowingly hold the opposite belief. If anything constrains what one can believe at will, it is what else one knowingly believes. My account gets around these difficulties by simply denying that the self-deceptive belief is formed intentionally. Instead, I claim that the belief is formed unconsciously via the operation of a mechanism which is closely analogous to those we already know about from the dual-processing literature.

In order to maintain that self-deception is nevertheless somehow an intentional phenomenon we need to identify that in the phenomenon for which the self-deceiver is responsible. We also need make sure to do it in such a way that preserves the idea that self-deception is a kind of motivated irrationality and that the self-deceived subject is one who manifests epistemic vice. My view locates the manifestation of this vice after the self-deceptive belief has been formed, in the subject's subsequent motivated failure to overthrow it. (Often this failure will require the subject to do, for motivated reasons, precisely nothing. Hence the name 'Self-deception as Omission'.) Self-deception retains its status as motivated irrationality — with a person-level desiderative state playing a crucial role — because the subpersonal mechanism which produces the belief does not alone suffice to explain why the belief persists

despite the abundance of evidence speaking against it.

The second half of Chapter 3 is devoted to exploring and critiquing rival views of self-deception. The discussion is framed by the following diagnosis. What makes the psychological dynamics of self-deception seem puzzling is the coming together of three philosophically important ideas: belief, intentional action, and the psychological unity of the self. Intuitively, self-deception seems puzzling because it is tempting to think that it involves acting intentionally so as to cause oneself — as a unified subject — to believe something one already takes to be false. Consequently, one might think that the air of paradoxicality could be removed by tweaking or downgrading one of the three crucial notions that play a role in that description. And this is precisely how I classify the rival views that exist in the literature.

One class of views says that the self-deceptive state doesn't involve full-blooded belief. The basic thought here is: what one can *believe* as a result of intentional action is constrained by what else one knowingly believes, but the same needn't be true of a state that is functionally similar to belief — it simply might not be subject to the same constraints. Another class of views puts pressure on the way in which self-deception is thought to involve intentional action. What seems puzzling about self-deception, according to this view, is that one seems required to intend *to deceive oneself*. But, theorists of this stripe will insist, that isn't actually required for self-deception to occur. The third class of views relieves the pressure found in the intuitive description of self-deception by denying the psychological unity of the subject. According to views of this type, one *part* of a self-deceived subject can act intentionally so as to deceive another part without paradox because the constraints that apply within a single psychological subject do not hold across subjects — as with interpersonal deception — or across parts of subjects.

Against all three types of competing views I argue, quite simply, that the two desiderata with which we began are not satisfied by them. Either they fail to truly resolve the dynamical difficulties, or they get the facts about responsibility wrong.

The most popular form that the first type of view has taken is to think of the self-deceptive state as a kind of elaborate pretense state. When one pretends that $p$ one can display many of the superficial features involved in believing $p$ — reporting it, acting as if it were the case, etc. — without being subject to the same constraints. The hope for such views is that pretending that $p$ (in a sufficiently complex way) will account for the self-deceptive syndrome in a way which is not inconsistent with knowingly believing that not-$p$. I argue that this is not a satisfactory way of dealing with the dynamical problem of self-deception. One can no more pretend at will — when the pretense is for the purpose of achieving some end that requires it to be concealed as pretense — than one can believe at will in order to achieve similar ends. Moreover, I argue that when the pretense is as elaborate as it seems to need to be to capture the self-deceptive syndrome, the distinction between pretense and belief begins to break down.

The relationship between my view and views of the second kind is slightly more subtle. Indeed, my view bears some close similarities to the most popular of them (Mele 1997). According to Mele's view, there is something that a self-deceiving agent can do intentionally that falls short of intending to deceive herself. With this much, I agree. However, I argue

that the particular mechanism that Mele appeals to leads him to a version of the dynamical problem of self-deception in at least some cases. In response to this I argue that we should instead embrace Self-deception as Omission, which makes no commitment to such a problematic mechanism. What is right about views of this type is that the self-deceiving act can't be a clear-eyed act of intentionally deceiving oneself, but Self-deception as Omission captures this feature of the phenomenon more effectively.

Finally, the views which claim that in order to make sense of self-deception we must acknowledge that the self is somehow fragmented are also right in a certain way. Indeed, the picture of the mind which emerges from cognitive science already commits us to a certain version of the fragmented self. The self has parts insofar as there are autonomous subsystems which comprise it and whose operations are unknown to the conscious subject. I take this opportunity to discuss the connection between the Freudian-style partitioning that inspired some accounts of this type (Pears, Davidson) and the sort of partitioning that dual-process theory commits us to. In the history of philosophy, some (e.g., Sartre) have been suspicious of the very idea that something properly mental could be unconscious. Sartre critiqued the Freudian idea of the unconscious on these grounds and I spend some time discussing Sartre's objections. Being skeptical of unconscious processes as such strikes us as quaint nowadays, and the correct response to Sartre clearly vindicates the mechanisms and modules of modern cognitive science. Still, the vindication is not trivial. Our subparts are indeed parts of us, and how we relate to them turns out to be of major significance for whether we act well, respond to reasons, are rational, and so on. In self-deception, however, that relation comes to the fore in the way that the conscious rational subject interfaces with the cognitive products of her subparts. This is what Self-deception as Omission captures by identifying the culpable failure in self-deception as a failure to judiciously check the deliverances of one's unconscious processing. Insofar as fragmentation strategies embrace the two-part structure of Self-deception as Omission, I have no quarrel with them. But historically with theories of this kind the crucial node where I want to locate the self-deceiver's intentional agency has been missing.

Chapter 4 is naturally connected with Chapter 3 in that I want to use what I have shown about self-deception and responsibility to probe whether any delusional subjects are responsible, in some sense, for their delusions. I propose to connect these two concerns by simply asking whether there is any overlap between self-deception and delusion. Somewhat naïvely, I argue that there is some overlap by adducing a case where this seems to be so, and suggesting that there could be others. However, I begin with a slight revision — to wit, a logical weakening — of Self-deception as Omission into what I call 'Self-deception as Omission*'. Self-deception as Omission required that the self-deceptive belief come about as the result of the operation a particular sort of subpersonal mechanism. Responsibility for self-deception was then located in a motivated failure to overthrow that belief. But does it matter how the self-deceptive belief was *formed*? It might not. According to Self-deception as Omission*, one is self-deceived if, however one's belief that $p$ was formed, the persistence of one's belief that $p$ is accounted for by one's motivated mismanagement of the evidence for and against $p$. That is, one counts as self-deceived if one fails to recognize or appreciate the

externally available evidence against $p$ because of one's desire that $p$ be true.

Moving to Self-deception as Omission* allows us to become clearer about the way in which self-deceivers are responsible. In order to do this, and in order to be clear about the kind of responsibility I think is at play in some delusional subjects, I distinguish between two kinds of responsibility. What I will call 'attributability' is a logically necessary condition on what I will call 'blameworthiness'. Some states or actions are attributable to me in the sense that I am the author of them or that they demonstrate a certain feature of my character. Nevertheless, I may not be blameworthy for performing those actions or being in those states because I may have an excuse. For example, suppose I have been tasked with reading and understanding one hundred unfamiliar pages of Leibniz in fifteen minutes. Failing to do so is something that would be attributable to me. It is something that I did, that demonstrates certain *ways that I am* — to wit, my limitations — but it is not appropriate to blame me for failing at this task. I have an excuse: it was too demanding. Self-deception seems to be essentially attributable. That's part of what we mean when we say that something is self-deception rather than some other kind of irrationality. But it is typically also blameworthy. The epistemic vice that self-deceivers manifest does not typically have an excuse. I argue that while there is overlap between self-deception and delusion — some delusional subjects satisfy Self-deception as Omission* — it will also typically be the case that they *do* have excuses. Their delusions are attributable to them but they are not blameworthy for them. This is an important result because it reminds us that we should think as much as we possibly can of delusional subjects as belonging within the moral community and not completely exempt — at least not *as such* — from responsibility attribution altogether.

The final chapter is about addition. The dialectic in this chapter it somewhat complex, but the basic thrust is nevertheless quite simple: thinking about responsibility in addiction must be done carefully and should be sensitive to the complexities of the phenomenon. In particular, we should not expect a clear, precise, or perfectly general answer to the question of whether and to what extent addicts are morally responsible agents. Working within the extended reasons-responsiveness theory from Chapter 2 allows us to do justice to this complexity. I begin by distinguishing between two different capacities for reasons-responsiveness that might be impaired in addiction, which I call 'recognition' and 'reactivity'. The remainder of the chapter is an investigation into the ways in which those capacities are impaired by addicts of different kinds and whether and in what way such an impairment could constitute a mitigating excuse.

One of the major goals of Chapter 5 is to argue that addiction should not be thought of as a condition which literally compels action. I am convinced that this is correct by testimonial and epidemiological evidence, and by scientific work on the effect of providing a positive stimulating environment for otherwise addiction-prone creatures (Alexander 2010). Still, room must be made in philosophical theorizing about action for thinking this is the case. Certain philosophical pictures of action make addiction seem inevitably like a compulsive phenomenon and they must be resisted if we are to make sense of what is going on. To that end, a large part of Chapter 5 is devoted to arguing against so-called 'Humeanism' about action, according to which action can only be motivated by desires. I argue that this picture

is based on a distortion of the idea of a desire and that it cannot account for the phenomenon of bringing oneself to act in response to an evaluative judgement. In order to do this, I argue that we must introduce *intentions* into our basic moral psychology.

Once we have introduced the idea of intentions, the impairments that addicts might face can thus be understood as (i) the inability to form the proper intentions and (ii) the inability to follow through on the intentions that one has already formed. I argue that failure of the first kind can be due to self-deception, and that if it is, it is not excused. However, failures of the second kind may be excused due to the particular character of addictive impulses and how they deplete one's willpower. The idea of willpower has recently been understood and operationalized as a System-2 capacity that people have in different quantities and which is depleted by effortful tasks, especially by tasks requiring self-control. I appeal to this recent empirical work, and to a certain conception of addictive desire, to argue that for those in whom addictive impulses have become deeply rooted their very persistence and unresponsiveness to rational judgement may constitute a gradable excusing obstacle to acting well.

The unity of what is in these pages is to be found in their concern with the lived practical reality within which questions about moral responsibility naturally find their home. The 'fringe' phenomena with which I have concerned myself — hardly examples of human agency at its best — are ideal candidates on which to carry out an investigation of this kind because, as non-ideal as they are, they are fiendishly common, and questions about how we are to think and feel about people who are in their midst unavoidably press themselves upon us. We may not be able to expect to find clear and precise answers to such questions, but that should not deter us from facing them. That is what I have tried to do in what follows.

# Chapter 2

# Skepticism

*Here's a hand.*

G. E. Moore

## 2.1 Introduction

It is natural to think that there is a close connection between conscious deliberative reflection, on the one hand, and agency and moral responsibility, on the other. Some of the reasons for this are based in perfectly ordinary reflection on phenomenology: the *experience* of agency is closely connected to making one's mind up about what to do. Still other reasons arise from minimal philosophical reflection: An episode of reasoning terminating in a decision to act in a certain way, followed by the performance of that very action, will not only be phenomenologically very salient to the agent, it also seems (perhaps rightly) to be a paradigm example of human agency. What could be a sounder basis for ascribing that action to that agent than the fact that the agent himself considered the reasons for and against performing it, evaluated their respective weights and merits, and himself took the decision to go through with it?

Not only do we take rational deliberation as sometimes sufficient for agency, we often think it is necessary: we often withhold ascriptions of agency and responsibility from creatures which lack sufficiently sophisticated deliberative and reflective capacities, such as children and non-human animals; we also often think that otherwise morally competent adult human beings may be 'off the hook' for decisions taken in circumstances where their reflective and deliberative capacities have been impaired by factors beyond their control — if I slip mescaline into your kale smoothie without your knowledge and you go on to cause harm to yourself or others, I am at fault, not you.

It is perhaps unsurprising that we *think* conscious reflection is where all the action is. After all, the conscious reflection that we engage in is the most conspicuous of our mental activities. It is among the mental activities that we are most vividly aware of happening, and

even among those — compare, e.g., pain — it is the one we are most inclined to think that *we* are in charge of. Those of us with any exposure to the Western philosophical tradition are made still more likely to draw this connection by the weight of tradition: thinkers as different and as vastly separated by time as Aristotle, Descartes, Kant, and many others have held that what distinguishes humans from other elements in the natural world is our rational capacities. Insofar as we are practically and not merely theoretically inclined creatures this involves also the capacity for rational self-determination in the practical sphere. And typically this is thought of as involving a kind of conscious reasoning.

Perhaps unlike the ethical views which have been traditionally allied to it, this conception of the human mind and what is most distinctive of it is open to empirical falsification. Maybe it really is purely a matter for debate from the armchair, e.g., what the *value* of a rationally-governed life is, or the extent to which one has a *duty* to cultivate one's rational capacities. However, the question of whether and to what extent we are actually rationally governed seems to be a straightforwardly empirical matter. It might just turn out that, according to our best science, human thought and action is not, for the most part, determined, or even guided, by (even imperfectly deployed) rational capacities. It might just turn out that, phenomenological salience notwithstanding, those effortful episodes of grueling practical ratiocination are not doing the work that we think they are in determining how we act and why. If this turned out to be the case, what would the implications be for our thinking about human agency and responsibility? Would our ethical thinking remain comfortably insulated?

In this chapter I will do the following. First, I will review the empirical results which I think are most relevant to philosophical thinking about morally responsible agency. These results have only recently come to be understood in a fairly theoretically unified way, and in order to make them bear on philosophical questions about responsibility as directly and forcefully as possible, some remarks on how to understand this theoretical unity will be required. Then I will proceed to articulate a skeptical position about moral responsibility based on the empirical results gathered. This part of the discussion will be heavily indebted to John Doris' recent *Talking to Ourselves* (2015). Doris himself sets up such a skeptical challenge as a way of motivating his own 'collaborativist' view of responsible agency. To a certain extent I will be following him in this. I think the empirical results do motivate a certain view of agency, but my view differs from his. I also think that the considerations Doris has in mind do not quite have the force that he takes them to have against the views of agency that he targets. My response will thus be twofold: (i) First, I will offer a reply on behalf of the theorists that Doris has in his sights. However, (ii) I will not ultimately be siding with those I defend from Doris and will instead conclude by arguing that a reasons-responsiveness theory of responsible agency does an even better job in the face of the empirical results discussed.

## 2.2 The Science

### (Somewhat Potted) History: From Dual-Processes to Dual-Systems

Since the 1970s, researchers in cognitive and social psychology have posited that aspects of human psychology as varied as deductive reasoning, social judgment, decision making, learning, and memory involve two importantly, and perhaps, fundamentally, different kinds of processing. The details of any given proposal tailored to a specific phenomenon may vary, but quite typically, one class of processes is characterized as being 'fast, effortless, automatic, nonconscious, inflexible, heavily contextualized, and undemanding of working memory, and the other as slow, effortful, controlled, conscious, flexible, decontextualized, and demanding of working memory' (Frankish and Evans 2009, 1).

One of the most striking features of the development of this theoretical perspective is that individual so-called 'dual-process' theories developed largely independently of one another in many different areas of psychology. This independent convergence is a very strong theoretical consideration in favour of the general outlook: researchers with very different concerns found themselves in need of a certain kind of novel mechanism or process to explain otherwise puzzling results that they were encountering. What began to emerge after several decades of work was that both the types of processes that the different researchers posited, as well as the processes that they were contrasted with, share clusters of features. Many psychologists now think that the mind is composed of (at least) two distinct *systems* comprised of collections of the two different types of processes. In this section I will review some of the earlier independent findings in which dual-process-type explanations figured prominently and float a proposal for understanding how they best hang together in a theory of cognitive systems.

#### The explanatory pattern

The idea that cognition (or at least, certain types of cognition) may involve two kinds of processes is not exactly new. As far back as 1960, for example, Jerome Bruner distinguished between what he called 'analytic' and 'intuitive' thinking (Bruner, 1960, 57-58, emphasis mine):

> Analytic thinking characteristically proceeds a step at a time. Steps are explicit and usually can be accurately reported...Such thinking proceeds with relatively full *awareness* of the information and operations involved. It may involve careful and deductive reasoning, often using mathematics or logic and an explicit plan of attack. Or it may involve a step-by-step process of induction and experiment...
>
> Intuitive thinking characteristically does not advance in careful, well-planned steps. Indeed, it tends to involve manoeuvres based seemingly on an implicit perception of the total problem. The thinker arrives at an answer, which may be right or wrong, *with little if any awareness* of the process by which he reached it.

Bruner here notes the difference between processes the operation of which one is aware and processes the operation of which one is not aware. There is a related distinction between so-called 'controlled' and 'automatic' processing that is also arguably at the root of a great deal of dual-process theorizing in psychology. Walter Schneider and Richard Shiffrin parse the difference between the two as follows (1977b, 127):

> Controlled [processing] is highly demanding of attentional capacity, is usually serial in nature with a limited comparison rate, is easily established, altered, and even reversed by the subject, and is strongly dependent on load. Automatic [processing] is...demanding of attention only when a target is presented, is parallel in nature, is difficult to alter, to ignore, or to suppress once learned, and is virtually unaffected by load.

Schneider and Shiffrin got at the difference between controlled and automatic processing in the lab by giving subjects a perceptual task (Schiffrin and Schneider 1977a). They presented their subjects with sets of four letters flashed rapidly on screen in front of them. The subjects' task was to detect whether target letters, which they are given beforehand, appear in each set. The target letters were chosen in one of two ways. Under one condition, designed to elicit controlled processing, the target letters changed comparatively frequently, after about 100 sets or so. They found that under this condition, if a second task requiring controlled processing was done at the same time, performance on both tasks declined. The second general condition was designed to elicit automatic processing. Under this condition, the target letters remained the same throughout testing, which lasted for several thousand trials. As one might expect, subjects became more adept as the trial went on and some even reported that the target seemed to 'jump out' from among the other letters. Furthermore, when observers did the controlled task and the automatic task at the same time, their performance was not affected. This would seem to show that that the automatically processed task runs along a separate cognitive 'track' and requires little or no working memory or conscious attention.

The essential differences between analytic and intuitive thinking on the one hand, and automatic and controlled processing on the other, were later combined by other researchers and applied in other areas. For example, researchers on memory and learning were led to distinguish between what they called 'implicit' and 'explicit' learning processes to account for subjects' performance on tasks such as the artificial grammar learning (AGL) paradigm. In this paradigm, subjects are exposed to a series of consonant strings generated according to a finite-state grammar, typically being told that they should try to memorize them to the best of their ability. Later, the subjects are told that the strings they were shown were rule-governed, and asked to evaluate novel strings for accordance with those rules. Subjects robustly perform above chance on these 'grammaticality' tests, leading researchers to hypothesize that there is a kind of learning which is not a consciously directed, deliberately controlled, process. Researcher Rebecca Gomez characterizes the difference between implicit and explicit learning this way (Gomez 1997, 154):

> Explicit learning is characterized as an active, voluntary, and purposeful process; one in which people generate and test hypotheses in order to adapt to changes in the environment. Such learning is accompanied by a high degree of awareness. Implicit learning, on the other hand, is characterized as a passive, involuntary process, one in which people "soak-up" complex, novel information with little or no awareness of the underlying structure or abstract rules.

Results such as these seem to show that learning is a more varied and complex phenomenon than was previously believed. But they show more than that. It's not just that there is a kind of learning, or an aspect of learning, that had previously gone unappreciated. It's that the newly discovered type of learning (like the 'new' kind of processing pointed to by Schneider and Shiffrin, and those that preceded them) had to be understood as possessing certain features, as operating in a particular way, in order to do any work helping to explain the experimental findings; on a single-process learning model, subjects' performance on the grammaticality tasks would be unexplained. So, researchers posited a mechanism whose operation could be invoked to fill in the explanatory gap, and could be understood to have been operative in the experimental settings in question. Subjects report having no awareness of the rules, so the learning must have occurred nonconsciously; subjects were not directed to attend to the patterns they were later able to identify, nor do they report having done so, so the learning must not have been voluntary, etc. This explanatory pattern repeated itself across many domains of empirical psychology.

Psychologists studying reasoning have also distinguished between two types of processes in order to explain their findings. For example, Peter Wason and Jonathan Evans (1975) noticed what seemed to be two discrepant findings on the famous Wason selection task (figure 1.1). On the one hand, the subjects seemed to be choosing cards in accordance with a 'matching bias' — that is, they responded by selecting cards which exhibit the features explicitly mentioned in the conditional statement.[1] (In our example, 'even' and 'blue'.) On the other hand, when asked about their choices, subjects produce rational-sounding explanations which make no mention of any such matching bias. Wason and Evans hypothesized that these seemingly conflicting results could be reconciled by recognizing that the process which causes subjects to exhibit the matching bias is distinct from the process that produces the introspective report; the former is unconscious and produces an intuitive solution to the abstract problem, and the latter is conscious and is meant to explain the rationales that the subjects reported for their choices.

A landmark study in a similar vein, done in 1977 by Richard Nisbett and Timothy DeCamp Wilson (Nisbett and Wilson 1977) purported to show the difference between belief — as a determinant of behaviour — and what subjects report as their reasons for acting. Nisbett and Wilson conducted their study in response to the suspicion, already shared by a number of psychologists that preceded them, that people may often have less than perfectly reliable access to their own higher-level cognitive processes. This suspicion was held, for

[1] This was verified by noticing that subjects still tended to choose cards displaying the features mentioned even if preceded by a 'not' in either the antecedent or the consequent (Wason and Evans 1975, 142).

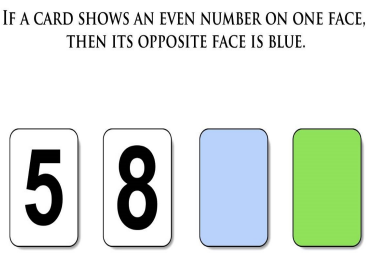



Figure 2.1: The Wason selection task: Which cards need to be turned over to evaluate this claim?

example, by George Mandler, who, writing before the Nisbett and Wilson study, summarized what he took to have been demonstrated up to that point (quoted in Nisbett and Wilson 1977, 232, emphasis mine).

> The analysis of situations and appraisal of the environment...goes on mainly at the nonconscious level...There are many systems that cannot be brought into consciousness and probably most systems that analyze the environment in the first place have that characteristic. In most of these cases, only the *products* of of cognitive and mental activities are available to consciousness...[U]nconscious processes...include those that are not available to conscious experience, be they feature analyzers, deep syntactic structures, affective appraisals, computational processes, language production systems, action systems of many kinds.

It seems perfectly ordinary to us now that things like deep grammatical knowledge and so much perceptual whirring and grinding should be operating below the level of conscious awareness. Perhaps this is due at least in part to the success of this, and related psychological research programs. But Nisbett and Wilson did not just want to show that there a lot of subpersonal processing and computation going which is below the level of awareness. They also wanted to show that some of those processes are driving behaviour at the personal level in ways that agents seem to be unaware of. The most famous of their results demonstrated an 'ordering-effect' on consumer choices. Posing as a market researcher, experimenters asked consumers which of a set of qualitatively very similar products they preferred and to explain their choices. As a matter of fact, there was a significant left-to-right position effect: subjects preferred articles positioned further to the right on the racks or shelves. But in giving their explanations 'no subject mentioned spontaneously the position of the article in the array' (Nisbett and Wilson 1977, 233-234).

Like the other researchers facing analogous results in their own domains, Nisbett and Wilson offer an *explanation* for the range of results they discuss. They propose that rather than consulting a (most likely unavailable) memory of the operation of some cognitive process to explain an experimental effect, subjects apply or generate causal 'theories' about the likely

effects of experimental manipulations. Nisbett and Wilson take this to imply the existence of two distinct streams of cognitive processes: one is responsible for the observable experimental effects, e.g., the left-to-right position effect. The other is responsible for the types of reports that people actually produce. The *distinctiveness* of the two processes, and the fact that they don't 'talk' to each other is what explains the pessimistic body of results concerning the divergence between what people report about what is going on with them, and what experimenters are able to verify is actually the case. Once again, facing their results drove researchers to posit independent processes.

Researchers have also discovered more troubling subterranean influences on thought and behaviour. In particular, so-called 'implicit biases' have also generated a lot of discussion among philosophers recently. Implicit biases are basically learned, automatically activated associations, whose influence on behaviour can be tested in the lab. Their influence needn't always be malign, but often it is, negatively affecting individuals' evaluations and judgements of, and interactions with, individuals in stereotyped or stigmatized social groups. The effects can be relatively minor, such as increased eye blinks when interacting with members of a stereotyped group. But the effects can also be grave. If a law enforcement officer is quicker to associate a black man with aggression and with dangerous objects, this increases the likelihood of misjudging his reaching for an ambiguous object as a threatening gesture. What is perhaps most striking about implicit biases is how pervasive they are, and yet how incongruent they are with our reflectively endorsed beliefs and values: even though the vast majority of us openly and sincerely express discomfort with stereotyping and endorse egalitarian values, the research shows virtually everyone harbours such biases. What this shows is that the force of all of this invisible cognition is not only potentially malign, it extends all the way up to fairly direct control over judgment and behaviour.

Although the distinctions between types of processing that have been drawn by different researchers are related, they are, of course, also importantly different. Figure 1.2 summarizes the sets of features that different researchers have contrasted.

This has become the 'standard menu'[2] of distinctions between features of what I shall call 'type-1' and 'type-2' cognitive processes. These are, of course, all very different features, but as even our cursory review of the research shows, they tend to co-vary, and as the successes of the individual dual-process research programs proliferated, some theorists began to ask whether the type-1 and the type-2 processes might be organized into complex cognitive *systems* which might explain this co-variation. Some theorists who are fans of this approach call the systems 'System 1' and 'System 2'.[3] Next, I would like to float a proposal for understanding what the unity of the two systems might consist in, and to provide something by way of a very simple argument for thinking that we should think there is this level of organization in cognitive architecture.

[2]This is the term used by Richard Samuels (2009).

[3]Thus there is a distinction between what we might call 'Dual-System Theory' and this or that (particular) dual-process theory (of learning, social cognition, etc.). The term 'Dual-Process Theory', however, has come to be used as an umbrella term standing for both the theory of systems and the class of individual dual-process theories. I will mainly use the term 'Dual-Process Theory' to refer to the former.

| **Type-1** | **Type-2** |
|---|---|
| Evolutionarily old | Evolutionarily recent |
| Unconscious, preconcious | Conscious |
| Shared with other animals | Unique to humans |
| Implicit knowledge | Explicit knowledge |
| Automatic | Controlled |
| Fast | Slow |
| Parallel | Sequential |
| High capacity | Low capacity |
| Intuitive | Reflective |
| Contextualized | Abstract |
| Pragmatic | Logical |
| Associative | Rule-based |
| Independent of general intelligence | Linked to general intelligence |

One very natural way of getting at whether there is this additional layer of cognitive organization is to ask whether we should think that there is an underlying 'natural kind' responsible for the apparent co-variation of the type-1 and type-2 features, respectively. One position is of course to deny that there is any meaningful co-variation to be explained. But this seems to make a mystery out of how so many different researchers working in such different areas could have independently come to find that they had need for so many of the same types of processes with their respectively clustered features. Why then would all of the type-1 processes seem to be unavailable to conscious awareness? Why do they all seem to be less affected by cognitive load? Why do they all seem to work associatively? Why are they fast compared with typical type-2 processes?

There are however two other possible sources of resistance to the suggestion that type-1 and type-2 processes — and the systems that they respectively constitute — are anything like natural kinds. The first would be to deny that there is any particularly deep connection between the type-1 features (or the type-2 features) taken as a class. The second would be to deny that there is any particularly deep difference between the two different kinds of processes.

According to the first suggestion, there is nothing that should make us think that there is any similarity in kind between a process' being, say, conscious and being controlled; or between being fast and intuitive; or between being rule-based and sequential, etc. On the face of it, this seems also to fly in the face of even the very brief and sketchy history of discovery outlined above. Many of the individual processes that various teams of researchers found that they were justified in positing simply did have more than one of type-1 features listed in the standard menu. It simply is true that the type-1 features tend to cluster together, and the type-2 features tend to cluster together.

According to the second suggestion, there is a kind of clustering amongst the type-1 and type-2 processes respectively, but there is no 'deep' or particularly interesting difference

between the two types of processes as a class. On this way of going, some processes may be faster than others, or more contextualized than abstract, and where they are, they should also be expected to be more associative than rule-based, more pragmatic than logical, etc. Now, of course it is true that some processes are faster, or more associative than others, but if this is all that a dual-process theory is claiming, it risks not being a very interesting theory; once we were ready to admit that there was so much as a difference between a faster and a slower process, we became thereby willing also to admit that some processes are faster than others. Further, it is very tempting to think that whatever it is that is the likely cause of the clustering of type-1 features together in a single process is also responsible for that process *not* possessing type-2 features, and by denying that there is any difference of kind between the two types of processes, this approach makes that thought unavailable. If the difference between type-1 and type-2 features is merely one of degree, it seems that there is nothing to rule out a process which, even though it groups together features which are, for the most part, more type-1-like, is also firmly in possession of a type-2 feature. But this is not what we find. What is so intriguing about the dual-processing results is that the type-1 and type-2 processes often *don't talk to each other*, and the possibility that there is no deep difference between the features that make a process of one kind or the other seems in tension with this kind of independence.

I therefore propose that we think of the difference between the two types of features (and the processes that possess them) as a difference between *natural kinds*. It is worth pausing to note that while I think the best interpretation of dual-process theory comes with this 'realist' commitment concerning natural kinds, this should not be mistaken for a commitment to those kinds being sets of identical members, having perfectly sharp boundaries, or being *entirely* independent of our goal-oriented practices. That is, not every type-1 process need exhibit every type-1 feature; nor should we expect that the difference between, say, a contextualized and an abstract process be an absolutely precise matter; nor should we think sense need be made of the idea that the identification of this particular list of opposed features could occur outside of the explanatory aims of the sciences of the mind. All of this notwithstanding, I think we can maintain that there is a perfectly robust sense in which the co-variation of type-1 and type-2 features can be understood as the kind of variation exhibited by members of a natural kind. Let me elaborate.

Traditionally, natural kinds have been conceived as groups of entities that share an essence. That is, on this sort of view, the essence of a kind specifies the conjunction of properties that are individually necessary and jointly sufficient for kind membership.[4] 'Essentialism' has the advantage of being able to explain the connection between a kind's being natural, and its supporting induction: It is *because* water is essentially $H_2O$ that the behaviour of water is inductively projectible. But, there seem to be many examples of intuitively natural kinds (like biological species) which don't fit nicely into the essentialist's account. Not every member of a given biological species will share all of its features. Not

[4]I gloss over here the difference between the 'nominalist' essentialism of Locke (1690) and the 'realist' essentialism of Putnam and Kripke (1980).

only is there wide individual variation, there is phenotypic variation. Of course, no essentialist account will require every member of a kind to have *all* of their features in common, but biological species shade into one another in a way that makes any attempt to isolate the relevant features look unpromising. There is nothing which guarantees that two organisms that are sufficiently genotypically different to prevent interbreeding will have more (say) morphological similarities than two individual organism belonging to the same phenotype. Richard Boyd's (Boyd 1999) highly influential *property cluster* theory of natural kinds is designed to overcome this kind of difficulty while still making sense of how natural kinds might support induction. According to Boyd, natural kinds are not sets of entities sharing an essence, but are instead groups of entities bearing 'various degrees of causally supported resemblance' (144).

For Boyd, the mechanisms that maintain resemblance amongst the members of a biological species are *homeostatic*, in the sense that deviations from properties that typify the cluster are not likely to thrive and reproduce. But not all kinds of resemblance between kind members are supported by homeostatic mechanisms. For example, consider butane ($C_4H_{10}$). Butane is an organic compound. Everything that is butane shares certain properties: a molar mass, a melting point, a boiling point, a specific heat capacity, etc. The regularity with which these properties are possessed by things that are $C_4H_{10}$ is causally supported in the sense that those properties supervene on the physical properties of a molecule with that composition. The regular occurrence of the supervenient properties of butane, like its boiling point, is causally supported by the the causal profiles of that upon which those properties supervene. What makes instances of butane share the properties that they share is thus not a mechanism which makes deviation from the properties that typify the cluster less likely to persist — after all, deviations from the the molecular structure of butane might just as well persist, but not as butane — so the mechanism is not homeostatic, but butane is natural kind nonetheless.

The sense in which I propose that we understand the claim that type-1 and type-2 processes are natural kinds is intermediate between the sense in which biological special are natural kinds and the sense in which butane is a natural kind. Just as no biological species is composed of identical members, we shouldn't expect every type-1 process to share every type-1 feature. Like the resemblance between instances of butane, the resemblance within the respective types of cognitive processes is causally supported, just not homeostatically. However, it is not clear that the clustering of type-1 and type-2 features is causally supported in the same way that the supervenient properties of butane are causally supported. That would required that the neural processes upon which those features supervene *themselves* group into two natural classes, and this is an empirical claim in need of much greater investigation.

Luckily, we don't need to rely on an empirical claim of that strength to formulate the natural kinds hypothesis in the way that I will appeal to in what follows. We know that the types of processes *do* cluster together into two groups — that's what we should conclude from the stream of results we have just surveyed. It is, therefore, highly likely that there is some kind of causal regularity in nature which is responsible for that clustering. And this

is all we need in order to support the following kind of induction: the fact that a process possesses one or more of the type-1 features probabilifies the claim that it will also possess other type-1 features and makes it less likely to possess type-2 features. I will appeal to this kind of proabilification a number of times below.

This is the understanding of dual-process theory whose consequences for responsible agency I would like to now evaluate. Let us proceed.

## 2.3 Empirically-Based Skepticism about Agency

Recently, some philosophers have raised concerns that the picture of the human mind that is emerging from the dual-process research program, and research related to it, poses problems for some philosophical theories of responsible agency. In particular, John Doris' recent *Talking to Ourselves* (Doris 2015) is (in part) an attempt to articulate just what such a skeptical concern would look like. Doris takes it that the threat is serious enough to motivate a move away from the views of agency that he targets and towards his own view. Much of what follows is an attempt to engage with Doris on these issues, though I will focus mostly on the negative part of Doris' project and will have comparatively little to say about his own positive proposal.

According to Doris, much philosophical theorizing about responsible agency is committed to a thesis that he calls 'Reflectivism'. He defines it for us thusly (Doris 2015, 19):

**Reflectivism** The exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do.

There is an intuitive way in which this looks like a traditional philosophical commitment; the exaltation of reflective judgment and rational self-government has been an idée fixe of many a grand figure in the western philosophical tradition. But Doris does not tell us explicitly who his targets are. It is plausible, however, that he has in mind theorists like David Velleman and Christine Korsgaard. Korsgaard says (Korsgaard 2009, 160):

> What makes an action mine, in the special way that an action is mine, rather than something that just happens in me? That it issues from my constitution, rather than from some force at work within me; that it is expressive of a law I give to myself, rather than a law imposed upon me from without.

On the plausible assumption Korsgaard thinks that giving myself a law is something that I do consciously, Korsgaard seems to be committed to reflectivism.

At first blush, reflectivism seems to be an empirical thesis, and Doris seems to be thinking that the science shows it to be false, so any philosophical view about agency or responsibility which requires its truth should be rejected on pain of skepticism about morally responsible agency. But this raises two related issues. First, *does* the science show reflectivism to be false? This is going to depend, of course, on how we interpret the science, but also on

*how much* of human agency needs to be 'ordered by reflection and deliberation' in order for reflectivism to come out true. Second, it is not straightforward what the implications are for a philosophical theory which in some way 'relies on' the truth of an empirical claim which is in dispute. This seems to depend in large part on what the philosophical theory is a theory *of*, and, of course, crucially, on the way in which the theory 'relies on' the truth of the empirical claim in question.

These issues are critical for understanding precisely how Doris wants to run his challenge. It would be one thing if (perhaps *per impossibile*) science had shown that *no* episodes of human action or judgment were responsive *in any way* to *any* kind of conscious reflective deliberation. But science has clearly not shown that. At best, science has shown that many domains of thought and action are not as responsive to conscious deliberation as we might have been antecedently inclined to think. I do not suppose that Doris is making the mistake of interpreting the science in this wild and implausible way, but now it is less than perfectly clear precisely how it is supposed to create a problem for reflectivism. If we take for granted that science has shown that type-1 processing is much more pervasive than we might have otherwise thought, and not more than that, does this show that human agency is not 'ordered by reflection and deliberation'? Perhaps it shows that it is not *always* so ordered. But what if it is only so ordered sometimes? Does this show that reflectivism is false?

Presumably there is no perfectly clear answer to the question of how much thought and action must be under deliberative control for reflectivism to remain tenable. How scarce would episodes of genuine agency have to be before a hard-headed reflectivist should be expected to flinch? The 'how much and how often' question concerning the actual role of reflective deliberation in agency seems like a difficult one to settle, and if Doris' challenge hinges on it being settled on one side of a difficult-to-locate line in the sand, the challenge can start to look a little murky. Perhaps one way of reading the challenge is as claiming that *however much* reflectivist-style agency can be made consistent with empirical psychology, it won't be enough to recover morally responsible agency as a sufficiently robust phenomenon to do justice to the role we in fact think it plays in our ordinary normative practices. It threatens to cause responsible agency to become so rare it ceases to be recognizable at all. This seems to be what he has in mind here (Doris 2015, 65-65):

> We are now in a position to schematize the skeptical challenge. Where the causes of her cognition or behavior would not be recognized by the actor as reasons for that cognition or behavior, were she aware of these causes at the time of performance, these causes are defeaters. Where defeaters obtain, the exercise of agency does not obtain.

The problem is simply that there are too many defeaters. And we just keep discovering more and more. Presumably, if the natural kinds hypothesis is true, this will only proliferate defeaters further. That hypothesis, if true, supports inductions of the form 'observing cognitive process P to have type-1(2) feature A(C), probabilifies the claim that P has type-1(2) feature B(D)', so *any* process which possess *any* type-1 features is more likely to possess the

deliberative-agency-impairing ones (unavailability to conscious reflection, automaticity) as well. Perhaps the purview of rational self-determination really has been peeled back too far for reflectivism to be the right picture of the mind to ground a theory of responsible agency.

The reflectivist could, of course, be quite hard-headed at this point. She could just put her foot down and claim that agency is an achievement and that we should not balk at its scarcity. However, I see two problems with this response. First, even if it were satisfactory in every other way, it hinges on a certain eventual answer to the 'how much and how often' question, one which we cannot simply assume. That is, presumably this response would not be convincing if the answer turned out to be 'almost none and almost never'. Further, (and Doris notes as much) even with what we know currently about the prevalence of type-1 processing, this would still be too much of a capitulation to skepticism. As we noted above, there seem to be defeaters lurking around every corner.

The reason the preceding line of response amounts to a capitulation to skepticism is because we are thinking of individual episodes of thought and action as either surpassing or failing to surpass some threshold determined by the distinction between type-1 and type-2 processing (how much of each sort is involved, which individual features are in play, etc.), the surpassing of which suffices to confer agentic[5] status on those episodes of thought or action. Then we are forced to ask: how much thought and action can we find on the agency side of the threshold? If the answer is 'not enough' we have given too much over to the skeptic. But there is another way of conceiving of what is required in respect of determination by reflection for thought and action to count as agentic. Science may have shown that the ordering of human life by deliberative reflection (or the causing of cognition and behavior by forces that the agent would recognize as reasons) occurs (effectively) much less frequently than we might have previously thought. But one might nevertheless insist that we can helpfully appeal to the reflectively determined life as a *regulative ideal*. On this way of thinking about the relation between reflection, on the one hand, and thought and action on the other, it is not the case that there is some threshold separating reflectively determined thought and action from thought and action that fails to be reflectively determined, with the latter kind failing to count to any degree as agentic. Rather, plenty of behaviour will count as agentic to the extent that it is regulated by the ideal of full or complete reflective determination.

This is a promising suggestion, but there are cross-cutting ambiguities hidden in the formulation thus far. What we might call 'normative reflectivism' might be a claim about what makes an instance of thought or action an instance of agency, or it might be a claim about what makes a creature an *agent*.[6] On the former interpretation, an individual episode

[5]My focus is obviously on morally responsible agency and not on the broader phenomenon of agency in general, but for ease I will use the term 'agentic' in what follows to stand as shorthand for 'morally-responsible-agentic'.

[6]Since it appeals to an ideal, one might think that a parallel ambiguity applies for the ideal: that ideal might itself either be the ideal of fully reflectively determined action on the one hand, or the fully reflectively determined agent, on the other. But it wouldn't make a whole lot of sense to try to compare either (i) an agent as a whole to an ideal of a full reflectively determined action or (ii) an individual episode of thought or behaviour to the normative reflectivist's ideal agent. Since what makes an agent an agent,

of thought or action counts as agentic if it is well-regulated by or stands in an appropriate relation of approximation to thought or action that is fully reflectively determined, as a matter of degree. On the latter interpretation, what is up for assessment is creatures, and we determine whether they count in general as agents by asking if *they* are well-regulated by or stand in some appropriate relation of approximation to the ideal agent, as a matter of degree. Which of these ideas should we take the normative reflectivist to be committed to? I think normative reflectivism ought to be understood *in the first instance* as a theory about agentic episodes, not what makes for agents. To see why, I think it will be helpful to look at a different philosophical position which makes productive appeal to the idea of a reflective ideal: interpretationism in philosophy of mind.

Interpretationists about intentional attitudes like Donald Davidson and Daniel Dennett think that we can learn something important about the mental by reflecting on the nature of interpretation. That is, they think that there is an important connection between being interpretable as a rational system and being truly describable as an intentional system. When we interpret someone, we assign propositional attitudes to him based on his behaviour and place those attitudes in a space of rationalizing relations with other bits of behaviour and other attitudes that we interpret him as having. This allows us to explain what they did and to predict what they might do in terms of reasons that he might himself be in a position to recognize.

Crucial to reason-giving explanations is the notion of rationality. In order to explain why someone did something in terms of their own reasons, there must be reason-giving connections between their various propositional attitudes, and between those attitudes and their actions.[7] When we interpret someone, we must therefore assume that they are largely rational; the ideal of rationality has a constitutive role in propositional attitude psychology because the idea of a reason-giving explanation is basic in this framework, and reason-giving explanations cannot proceed without the assumption of rationality. Assuming that the subject we are trying to interpret is rational is what allows us to move from a synchronic description of the state of a putatively intentional system to explanation and prediction of how that system has, or might, behave.

So interpretationists assign a very important role to rationality, which is partially analogous to the role that reflectivists assign to reflective determination. Interpretationists, one might think, are then faced with a similar problem to the problem of scarcity for reflectivists. What to make of the manifest prevalence of so much human irrationality? The typical reply is to claim that irrationality can only be understood as such as a failure of rationality against a backdrop of a more-or-less rationally unified set of attitudes. What makes an attitude an

on a reflectivist picture, is surely a matter of what sorts of thought and action she engages in, and how reflectively determined those episodes of thought and action are, it is unclear how being an agent could be (i) determined by a comparison to some particular action or (ii) what the comparison would be between an individual episode of thought and action and the reflectvist's ideal agent other than simply a comparison to a fully reflectively determined action.

[7]One is here tempted to say that rationality *just is* the obtaining of these reason-giving connections within networks of such attitudes.

irrational one is that it fails to cohere in the right way with the other attitudes that one holds which themselves do exhibit the required degree of coherence to constitute an intentional system.

Notice that this reply (however satisfying it ultimately is) relies on the idea that interpretationism is *in the first instance* a theory about what makes for intentional systems, not (in the first instance) what makes for some bit of intentional thought or action. That may well have implications for which bits of thought and action count as intentional given what commitments interpretationism has about how intentional states must be related to one another in an intentional system to count as such states in the first place. But this is a *consequence* of buying the interpretationist's picture of the mental, not an independent part of the theory.

The fact that the analogous line of reply does not seem available in the same way for a normative reflectivist seems to indicate that normative reflectivism is not in the first instance a theory about what makes something a morally responsible *agent*. The normative reflectivist can't say 'Ah, but you can't understand what it is for an episode of behaviour to fail to be agentic without first thinking that the creature in question is an agent by the lights of my theory.' This is because there is nothing about an episode of thought or action closely approximating or failing to closely approximate the ideal of complete reflective determination which requires that the agent him or herself approximate over his or her history the ideal of the reflectively determined agent. Indeed, the order of explanation would seem to go the other way around. *What it is* to approximate or fail to approximate the ideal of the reflective agent just is to have a history of individual episodes of thought and action that, taken together, do or do not approximate the ideal of reflectively determined thought and action.

This suggests that normative reflectivism is best understood in the first instance not as theory about what makes for morally responsible agents, but rather a theory about what makes for episodes of morally responsible agency. The worry about scarcity is thus a worry about episodes of agency and not about scarcity of agents themselves. However, it seems this kind of normative reflectivism needn't be overly troubled by the falsity (or limited purview) of (descriptive) reflectivism, the thesis that Doris discusses. As before, the correct interpretation of the science does not imply that there are no cases of effective deliberatively determined agency, but it is not obvious that the normative reflectivist requires even this much. So long as there are episodes of agency which approximate the (perhaps never realized) ideal of fully reflectively determined thought and action to a non-trivial degree, this stripe of normative reflectivist is not faced with claiming that agency is scarce — these approximations, after all, might not be scarce — and is certainly not committed to full-blown skepticism about agency.[8]

[8]The idea of appealing to normative reflectivism as a response to Doris' challenge I encountered first in a seminar the Manuel Vargas taught at Berkeley when he visited in Spring 2016. During a book symposium at the 2016 APA Pacific Division Meeting, Vargas also made a reply in his comments to Doris which appealed to the idea of normative reflectivism. Doris' response was to accuse Vargas of being an 'anti-realist' about agency. But I do not see that the position has this implication, for the reasons just given.

As far as a response to this first way of understanding Doris' challenge goes, I think appealing to a normative version of reflectivism does the job quite well. Although this version of reflectivism has some affinities with the view that I ultimately want to develop and defend, it is not my preferred view. However, the problem with the view is not that it is committed to skepticism about agency, but that it is committed to the wrong ideal of human thought and action. Reflection has a role to play in human agency, even a very important one, but there are ways of being an agent — that is, a creature with the capacity to recognize and respond to reasons — which do not consist in approximating the reflectivist ideal. But that is to get ahead of ourselves somewhat. Before coming to my preferred view, there is still another way of understanding Doris' challenge which has some force, and needs to be considered.

That other way is this. One might think that since the ascription of responsible agency is something that we *do*, and in some cases (perhaps very many) it can matter very much whether we do or don't do it, we should take great care to ensure that our confidence in such ascriptions is as high as we can reasonably strive to make it. When we ascribe responsible agency to someone, we are putting them on the hook for what they do, we open them up to all manner of normative assessment they would otherwise be insulated from and we risk treating them *unfairly* if we ascribe responsible agency where there is none. The empirical results should give us pause before we ascribe responsible agency in any given case. After all, we were inclined to think rationality was in the driver's seat all along, and we were wrong. And we are discovering every day just how wrong we were. Science hasn't pushed conscious deliberation out of the picture altogether, but it has pushed it back far enough that our confidence that it is doing any work in any given case should be fairly low. The natural kinds hypothesis would presumably help here as well, and in a similar way as before. Again, that hypothesis, if true, supports inductions of the form 'observing cognitive process P to have type-1(2) feature A(C), probabilifies the claim that P has type-1(2) feature B(D)', so *any* process which possesses *any* type-1 features is more likely to possess the deliberative-agency-impairing ones (unavailability to conscious reflection, automaticity) as well. Since there are a great many type-1 features, and we seem to be learning that more and more processes are in possession of type-1 features, our credence that any given process does not have agency-impairing type-1 features ought to be diminished.

Notice that this way of running the challenge combines a standard skeptical worry with a cautionary principle and a concern for fairness. The standard skeptical worries on their own invite the standard anti-skeptical replies. Just as one is tempted to say to the 'Cartesian' that it was never part of knowing that I need to know that I know, or that I need to know that certain far-out seeming possibilities incompatible with my knowledge do not in fact obtain, one is tempted to say that here I do not need to know that defeaters do not obtain, it simply must be the case that defeaters do not in fact obtain, and we are supposing that there are indeed cases like that (given modesty about what science has and has not shown). That is fine as far as it goes. But in this skeptical scenario, unlike in its standard Cartesian counterpart there is something at stake which is both important and, in an important sense *up to us*. According to the standard anti-Cartesian reply, it could be the case that I lack all

manner of ordinary empirical knowledge if the skeptical scenario in fact obtains. But so long as the scenario does not obtain, things are about as golden with me as I was antecedently inclined to take them to be. And that seems to be the end of the story. Everything is probably fine, and if it isn't that just reflects the actual poverty of my epistemic position, one which I wouldn't be in a position to know about anyway. Them's the breaks. But things are not this way with responsibility and agency. If we are wrong about whether people are in fact suitable targets for blame, we run the risk of treating them in a way is potentially seriously unfair. We can't just say 'if it turns out we are right, then everything's fine; if we are wrong, well them's the breaks.' If we are wrong we may be implicated in unjust treatment.[9]

For this reason, I think this version of the challenge can seem like it has some force. Moreover, it might seem that the appeal to normative reflectivism won't work here, at least not in the same way. Even if we were to move to a version of reflectivism which were normative in the way just elaborated, rather than descriptive, these considerations should still give us pause before ascribing (whatever degree of) morally responsible agency to someone. The version of normative reflectivism that takes reflectively determined action to be a regulative ideal works by assuming that we understand what that ideal consists in, and that we know how to make the relevant kind of comparison between that ideal and its more humble counterparts, which itself obviously involves having a good understanding of what is going in those cases. This way of running the challenge puts pressure on this second idea; *for all we know* there are agency-impairing (because non-reflective) processes at work all over the place. If that's right, then any comparison with the ideal might simply not be apt. But if no thought or action is meaningfully compared to the normative reflectivist's ideal, there would appear to be no thought or action that approximates that ideal to a non-trivial degree, and this suggests that there would be no morally responsible agency.

However, a little reflection shows that the normative reflectivist needn't capitulate this much to this second version of the challenge. This way of running the skeptical challenge attacks the justifiability of our practice of holding people responsible on cautionary grounds. But, now that we have seen that normative reflectivism is in the first instance a theory of episodes of agency, we can see normative reflectivists are in good position to deflect such a challenge as misguided. The move to normative reflectivism from descriptive reflectivism, in response to the first version of the challenge was, in effect, to say that some degree of agency is compatible with the presence of defeaters. If that's right, then a global challenge to agency looks like it misfires. 'We've got a theory of morally responsible agency, and if it's true, then there are episodes of agency, and we can be confident that there are, so we can meet the standard of caution.' — the equivalent, perhaps, in this domain of: 'Here's a hand, so I can know there is an external world...'

[9]The idea that it is a concern about fairness which underlies our concern to justify responsibility-based practices is developed extensively by Wallace (1994).

## 2.4 Reasons-Responsiveness, Default Action, Turning on System-2

In this section I want to ask how normative reflectivism, as elaborated above, compares to so-called 'reasons-responsiveness' views of responsible agency. If the preceding has been correct, then normative reflectivists can resist both versions of the skeptical challenge. However, I want to argue that a reasons-responsiveness view, suitably understood, can also resist the challenges, but that it should be preferred to a reflectivist view. The views do have some affinities, but my strategy will be, in effect, to argue that there is something which has been worth defending about reflectivism but that it is also, or better, captured by the reasons-responsiveness view. Further, what is right about reflectivism is not the whole story. The view is too narrow. There are important cases of responsible agency that it does not capture, or can only capture by effectively being assimilated into a kind of reasons-responsiveness view. First, a bit about the basic commitments and motivations for reasons-responsiveness theories of agency.[10]

According to a reasons-responsiveness theory of responsible agency, what makes individuals responsible agents is their possession of a general capacity (or a suite of capacities) to recognize the force of reasons for action and to act on the basis of those reasons. This corresponds in an intuitive way with our ordinary beliefs about what sorts of creatures are responsible agents. Normal adult human beings are responsible agents and have this (these) reasons-responsive capacity(ies) whereas young children and most non-human animals do not. Similarly, there are simple cases where we think that an agent is not responsible, where the reason we have this judgement seems to be that the person in question did not have the capacity to respond to reasons. Cases of mental impairment and manipulation are typically like this. Suppose I have been hypnotized in such a way that causes me to emit certain highly charged racial slurs upon hearing the sound of my doorbell. There are reasons why I should not emit those slurs — I may even in a certain sense recognize them — but my action is not responsive to them. My action is not responsive to reasons because the reasons to refrain seem to be making no difference to whether I perform the action or not; I am unable to get my recognition of those reasons to have the appropriate effect on my behaviour. Contrast this with cases where action issues from someone's ordinary faculties of practical reasoning. If I am deciding to go to the party or work through the evening I consider the importance of my goals (productivity vs. leisure) and the features of the options available to me. If the party promises to be fun enough I may forgo productivity in favour of it, but if I have received credible evidence that it is likely to be dull, I might reconsider. Or, I may come to recognize a new reason, such as an approaching deadline, that had slipped my mind,

[10]It is not clear whether or not Doris means for his challenge to be directed at reasons-responsiveness theorists. Part of this is due to his unwillingness to explicitly say who his targets are, and to make reflectivism, as he says 'a composite face which looks a little like many faces but not a lot like any particular face' (Doris 2015, 17). But at least one reviewer takes Doris' targets to include 'Korsgaard, Wallace, Darwall, and Velleman' (Shoemaker 2015). To the extent that Doris means his challenge to apply to views of this kind, this section can be read as a response.

that would also cause me to change my mind. The various mechanisms that are relevant to the production of my decision and my eventual actions are reasons-responsive because what reasons there are, and their relative force, has a role in determining which action I ultimately perform.

Reasons-responsiveness theories also handle Frankfurt-style cases well. In Frankfurt-style cases the agent is intuitively responsible even though he couldn't have done otherwise. In such cases there is an actual sequence of events that proceeds in a way that grounds the attribution of moral responsibility — e.g., Jones kills the Prime Minister because he believes him to be responsible for some injustice — even though external factors were in place to ensure that Jones would have killed the Prime Minister even if he hadn't wanted to — Black would have activated a mind-control chip in Jones' mind if Black had seen that Jones was not to go through with it on his own. The external factors make no difference in the actual sequence, but they do eliminate any alternative possibilities for Jones. However, the mechanism which causes Jones to act in the actual sequence is reasons-responsive. If Jones had recognized a reason for not killing the Prime Minister — perhaps he did not perpetrate the injustice that Jones believed him to, and he comes to see that this is so — his motivational state would have been altered. Unfortunately for Jones, he is in the grip of Black's counterfactual influence, so even if this were the case, he would have gone through with it anyway. As it happens, in the actual sequence Jones acts from a reasons-responsive mechanism and we judge him to be responsible; had he acted from the mechanism Black installed in his head we would have not judged him morally responsible, seemingly because his action would have issued from a mechanism which is *not* reasons-responsive.

Reasons-responsiveness theorists needn't deny that consciously recognizing reasons is a way of recognizing them, but it also needn't be the only way. Nor need it be the case that the capacity for recognizing and responding to reasons be in fact exercised in any given case for the agent to be responsible. Allow me to elaborate on both points.

## De Re Responsiveness

There is no need for a reasons-responsiveness theory to be restricted to the *conscious* recognition of reasons. The reason is this: There are cases where agents seem to act in a way which is morally praiseworthy or blameworthy *without* exercising their capacity for reasons-responsiveness in the way that a reflectivist would require them to do so. This is perhaps the right place to put a particular Strawsonian card on the table. As I said in Chapter 1, I take it that *in the first instance* a theory of morally responsible agency is a theory of the conditions under which agents are the appropriate targets for the reactive attitudes. Thus, if we want our philosophical theory to conform to the intuitive data about cases, and our theory is a theory of responsible agency, it should be guided by the intuitive 'data' concerning the conditions under which agents are morally praiseworthy or blameworthy. Thus, being faced with cases where agents seem to be praiseworthy or blameworthy but where it is implausible that they enjoyed anything approaching conscious awareness of their reasons for acting puts pressure on reflectivism. Some cases like this will be like the much-discussed

'Huck Finn' case. Huckleberry Finn is in a lot of ways a naïve, morally unsophisticated fellow. He has absorbed to a considerable extent the morality of his place and time. He thus professes, on a conscious level, the belief that helping his friend Jim escape would be tantamount to 'stealing' Miss Watsons' 'property'. Nevertheless, when the time comes, he helps Jim escape. Recently, Nomy Arpaly has compellingly discussed such cases. She says (Arpaly 2001, 76–77):

> There is [an interpretation of what Huck does where he] is morally praiseworthy for his action, and I would guess this is the scenario Mark Twain had in mind, though whether he did is of no consequence for my argument. On this interpretation, Huckleberry Finn is acting from neither squeamishness nor a desire to upset the adults. Rather, during the time he spends with Jim, Huckleberry undergoes a perceptual shift. Even before meeting Jim, the way Huckleberry viscerally experienced black people was inconsistent with his "official" racist views. [Huck] is a deliberative racist and viscerally more of an egalitarian. But this discrepancy between Huckleberry's conscious views and his unconscious, unconsidered views and actions widens during the time he spends with Jim. Talking to Jim about his hopes and fears and interacting with him extensively, Huckleberry constantly perceives data (never deliberated upon) that amount to the message that Jim is a person, just like him. Twain makes it very easy for Huckleberry to perceive the similarity between himself and Jim: the two are equally ignorant, share the same language and superstitions, and all in all it does not take the genius of John Stuart Mill to see that there is no particular reason to think of one of them as inferior to the other. While Huckleberry never reflects on these facts, they do prompt him to act toward Jim, more and more, in the same way he would have acted toward any other friend.

I agree with Arpaly that Huck seems praiseworthy.[11] Perhaps part of what undergirds our judgement that Huck is praiseworthy is that we think that, were he transposed to our place and time, Huck would not profess to have the beliefs that he explicitly claims to have. His commitment to the conventional morality of his milieu does not penetrate beyond the surface level. On the interpretation of the story that Arpaly favours, Twain is trying to get us to see the innocent humanity of Huck's simple ignorance. As Arpaly puts it, our response is to what we take to be Huck's 'deep moral concern'. It seems to us like, however ignorant he is of his the goings-on in his own mind, and however willing he is to parrot conventional wisdom, Huck is displaying a responsiveness and sensitivity to morally significant features of the situation. This kind of responsiveness, occurring as it does under no guise, has been dubbed '*de re* responsiveness'.[12]

[11] I can also agree that Huck might have been *more* praiseworthy if his action and his professed beliefs weren't so incongruent. This is presumably at least in part because we are responding to Huck's character, and his character would be a better one if it didn't harbour this incongruence.

[12] I may thus occasionally contrast responsiveness *de re* and responsiveness *de dicto*.

Cases like these highlight one thing that is lacking in a reflectivist-style understanding of agency. What is important when we are assessing the aptness-for-normative-assessment of agents is not which normative features they have passed through their conscious awareness, but which such features should be appealed to to give an intentionally sound explanation of their conduct. A hark back to the interpretationist theory of intentional attitudes discussed above might be helpful: what makes a bit of intentional psychology (including the attribution of sensitivity to a normative reason, it seems) appropriately attributable to an agent may be constrained by what makes for the most plausible explanation of what the agent did or thought, given the other things that the agent thinks or has done. From this perspective, whether such a reason passed through conscious awareness seems rather beside the point.

Perhaps reasons-responsiveness theories traditionally conceived have not countenanced this kind of sensitivity to reasons, and to the extent that that is correct, this suggestion is a revision or an extension of those theories. But I think reasons-responsiveness theorists should embrace this extension — especially if they are sympathetic to the broadly Strawsonian thought mentioned a little while back. If what we are primarily concerned with is the appropriateness of the reactive attitudes, and the data of intuition seem to indicate that those attitudes are applicable in cases where the agent had no conscious awareness of the reasons he was acting upon, *but nevertheless* it seems plausible to explain his conduct by appeal to sensitivity to those very reasons, then the applicability of those reactive attitudes will be grounded in a reasons-responsiveness theory of responsible agency according to which such attitudes find their correct application where agents are appropriately responsible to reasons.

It is important to note that I am not denying that part of what we are doing when we are responding to Huck is making a characterological assessment. I think we can distinguish, in general, between assessments of character, and judgments of praiseworthiness and blameworthiness. However, these kinds of judgments are often related and can partially overlap. It may be that part of what grounds our judgment of Huck's praiseworthiness is a characterological or aretaic judgment ('He has a good heart'; 'He means well deep down' etc.). But what is important for present purposes is just to notice that the judgement we make of Huck can be cashed out in terms of his responsiveness to the moral reasons that we think bear on the case. Our thinking that he has a good heart, or means well deep down may play a role in judging him praiseworthy, but that's because his having those features inclines us to read the case as one where the relevant moral reasons are playing a role in determining his action. Part of what it is to mean well is to exhibit a kind of sensitivity to such reasons, though some do it better than others, and some do it with conscious awareness, while others do it without knowing that is what they are doing, or even while holding contrary judgements consciously in mind.

## Indigent Responsibility

There are further cases where agents are appropriate targets of praise and blame where their capacities for reasons-responsiveness may even remain unexercised (at least with respect to

the reasons deemed relevant for the action or omission in question.) Many cases like this will be examples of culpable ignorance. Cases like these, and the kind of moral responsibility that attends them, will serve as a model for one aspect of the account of self-deception that I propose in the Chapter 3. Holly Smith's (1983) discussion of culpable ignorance opens with the case of a physician who treats a newborn's infant's respiratory distress by exposing her to high concentrations of oxygen, and so causes severe eye damage. There is no doubt that what the physician did, what he was causally responsible for, was bad. As to whether is blameworthy, however, it seems we must ask whether there is some sense in which he *ought to have known better.* If we suppose that there had never been any previously documented cases of such a treatment causing eye damage I think there isn't much of a temptation to think the doctor is blameworthy. On the other hand, if it well known by anyone who was paying attention in medical school that high concentrations of oxygen can be damaging to infants' eyes, it is pretty clear the doctor is blameworthy. Indeed, I strongly suspect there will be a spectrum of blameworthiness here and that it would enough to ground a (perhaps weaker) judgement of blameworthiness if the current issue of the Journal of Medicine reported on this very phenomenon, but that journal remained unopened on the doctor's desk because he chose to go golfing instead of reading it when it arrived last week.

Cases like these have the following general structure: an agent's failure to recognize and/or respond to reasons is caused by a failure to do what is necessary in order to apprehend or be moved by reasons. Because it is characterized by a failure to do something, I call this phenomenon 'indigent responsibility'.

This notion can be given some considerable teeth once we notice that whether an agent engages her System-2 processing or not, is often best understood in terms of whether the agent *chooses* to activate this form of processing. I have been downplaying the important of coming to be consciously aware of reasons for one to be rightly said to be acting in accordance with them, but this is perfectly compatible with the claim that there are some reasons that are best, or only, appreciated via the exercise of serial, reflective, System-2 processing. And if the agent's insensitivity to such reasons can be traced to a failure to engage System-2 which is analogous to the indigent physician's decision not to read the medical journal,[13] I think there is plenty of room to say, just as with the doctor, that the agent can be held responsible. To see why we should say that the engagement of System-2 is something which is under the agent's control, consider the following example from Kahneman (which we will return to more than once). Kaheneman says: 'Don't try to solve it, but listen to your intuition.' (Kahneman 2011, 44):

> A bat and a ball cost $1.10 The bat costs one dollar more than the ball. How much does the ball cost?

[13]The doctor's failure may in part be a failure to engage System-2, but it also may not be. If he chooses to nap instead of reading the journal he opts for cognitive ease over cognitive strain, but we can imagine that he left the journal unopened to play chess, and I think the sense he is blameworthy, to the extent that we have it, remains. His indigence in that case thus might be more of a professional rather a cognitive variety.

I myself experienced very strongly the presentation $0.10 as correct upon first encountering this example. It just seems to come unbidden. That answer is, of course, incorrect. But a little of further thought is required to see that this is so. And it seems to be thought of *a different kind*. For one thing, unlike the intuitive answer, one must put in a little bit of effort to see the correct one. Not much is required in this case, but I submit it is enough to be noticeable. If you don't *try* to do it, the answer simply doesn't come. Not so with the intuitive answer. It comes, it seems, whether you like it or not. This suggest that whether one engages in the further type of effortful System-2 thinking is largely a matter of choice or decision.

I think that the natural kinds hypothesis also helps us here. If it is correct, then the phenomenological difference between the two types of processing illustrated by the bat and the ball example can be given a causal underpinning. It's not just that the two types of processing seem different to us from the inside, that felt difference corresponds to a real difference in the sorts of things that are going on under the hood.

If this is right, then I think we should say that there are cases where one can be on the hook for *not* engaging System-2 where *had you* engaged System-2 you would be a position to easily come to appreciate some reasons that are otherwise unavailable to you.

Notice that cases of indigent responsibility are not just cases of derivative responsibility. In cases of derivative responsibility, the agent remains responsible for her action even though her reasons-responsive capacities are impaired because that impairment traces back to an action for which she is uncontroversially responsible, such as when someone voluntarily undertakes to drink too much and then gets in his car to drive. Cases of indigent responsibility don't (or needn't) involve any *impairment* to the agent's reasons-responsive capacities, but just a failure to exercise them. There may be partial overlap between these sorts of cases, however, such as if an agent becomes cognitively exhausted via her voluntary participation in an activity that leaves her less willing to undergo strenuous System-2 activity later on, where engagement of System-2 is necessary to appreciate a class of particularly pertinent reasons. (I will have considerably more to say about the willingness to exercise System-2 as a depletable resource in Chapter 5.)

## Default Action

We are now in a position to introduce an idea which will round out our review of the features and advantages of the reasons-responsiveness theory, but which will also be useful for us in later chapters. Restricting ourselves to causes that are, in some appropriate sense, internal to the agent, let a *default action* be an action (or omission) that an agent will perform (or

fail to perform) unless she intervenes via the activation of System-2.[14] [15] According to this definition, falling into the trap on the bat-and-ball problem is a default action. I will argue later that being (or, better: remaining) self-deceived and engaging in deeply entrenched addictive behaviours also count as default actions. A default action is not to be confused with the action that the agent is mostly likely to perform. After all, she may be diligent and it may be most likely that she kicks in System 2.

Whether an agent is responsible for performing a default action will depend on whether the agent's default mechanisms are reasons-responsive (that is, whether in virtue of the default mechanism's being hooked up in the right way, the agent is *de re* responsive to reasons) and whether the failure to activate System-2 is an example of indigent responsibility.

A reasons-responsiveness theory with all of these features handles both skeptical challenges in ways similar to the way normative reflectivism does. The first skpetical challenge works by claiming that defeaters are present in enough cases to imperil agency, or to make it unacceptably rare. But the defeaters are only defeaters if one's conception of agency is a descriptive reflectivist one. The reasons-responsiveness theory that I favour can avoid this problem by appealing to *de re* responsiveness and indigent responsibility. The second version of the skeptical challenge entreats us to make sure we are quite confident someone is an appropriate target for responsibility attribution and cautions us that the empirical evidence should give us pause. The Strawsonian move made above, coupled with an appeal to *de re* responsiveness and indigent responsibility will get us around this problem. Let me elaborate.

The extended reasons-responsiveness theory that I favour also deals with the second skeptical challenge in a similar way to normative reflectivism. The second skeptical challenge appears to be a global challenge to responsibility-attribution practices in the name of caution and fairness. Just as normative reflectivism, when properly understood, is seen to be, in the first instance, a theory of episodes of responsible agency, the extended reasons-responsiveness theory is also. The theory simply says that someone is responsible if and only if they are the appropriate target of (among other things) the reactive attitudes, and those attitudes are appropriate if the agent is responsive to reasons *de dicto*, responsive to reasons *de re*, or indigently responsibly. 'Here's a hand....'

One can imagine a variety of reasons-responsiveness theory which was extensionally equivalent to a descriptive reflectivist theory. This would be a theory that said that the only way one could be responsive to reasons was by reflecting on them. Indeed, a descriptive reflectivist of this sort would merely be helping herself to a description of what one is doing when one's

[14] The reason for the leading qualifier should be fairly clear: This definition says that a System 2 intervention is necessary for the prevention of default action, but clearly there are plenty of factors outside of the agent that could easily interrupt the flow of System 1 activity. For example, a severe enough earthquake probably suffices to interrupt pretty much any mental process, but the fact that an earthquake struck at the moment I first encountered the bat-and-ball problem and interrupted me before System 1 could yield an answer doesn't make it the case that the initial production of the wrong answer is not a default action.

[15] As defined, actions which are performed after an appropriate exercise of System-2 activation will count as default actions if the exercise of System-2 didn't make a difference to what action the agent performed but because of the connection with indigent responsibility we will be largely interested in cases where the exercise of System-2 would have been likely to make a difference.

decisions are ordered by conscious reflection, viz., one is taking stock of what reasons there are which bear on the decision. Whether this is the sort of view that reasons-responsiveness theorists have traditionally had in mind is difficult to say, but it is precisely the sort of view for which the empirical results seem to make real trouble. If consciously reflecting on reasons were the only way we could be responsive to them, and if the science has shown that we effectively do this a lot less than we might otherwise have thought, this would seem to show that we are effectively responsive to reasons far less frequently than we thought, and therefore that agency is much rarer than we thought. The move to normative reflectivism is a move to a position on which responsible agency is compatible with some defeaters because the presence of defeaters only shows that agency is imperfect relative to the ideal, and this does not mean it does not obtain. The extended reasons-responsiveness theory can deal with defeaters in a similar way. The basic parallel thought would be: defeaters are only defeaters (where they are) for *de dicto* responsiveness. Indigent responsibility is compatible with the presence of defeaters (if the right conditions are met), as is *de re* responsiveness.

Still, I think there is more to say, and that this will help us see why we should prefer the extended reasons-responsiveness account. For illustration, consider the following example:

> Colin is considering moving from Delaware to Colorado for work. This will require being further from his daughter, who he sees from time to time, but who is typically in the care of her mother. Colin's boss at his current job is black (Colin is not). Colin also has the following quirk that he employs to help him decide what to do. He feels that he can focus more on the facts that are relevant to the decision if he writes 'Colorado' and 'Delaware' down on a piece of paper, rapping his pen against the page while he ruminates. He typically writes them in that order, figuring that when other things are equal (which, attempting to be unbiased, he strives to make them as much as he can) alphabetic order is good as any.

Three potentially agency-impairing factors are at work in the case. (i) Colin may harbour implicit bias against his current boss, (ii) Colin may be subject to 'implicit egotism'[16] with respect to the name of the state he is considering moving to and (iii) Colin may be subject to an ordering effect when he ruminates by staring at the page with the names of the states written as they are.

Suppose Colin moves to Colorado, and this has an adverse effect on his relationship with his daughter, and subsequently an adverse effect on her well-being. Is Colin blameworthy for his decision? Let us suppose that the agency-impairing factors are defeaters in Doris' sense: together they cause him to take the decision, but he would not recognize any of the effects as reasons for moving to Colorado. I do not have a strong intuition one way or the

[16]Implicit egotism is at play when subjects unconsciously prefer things associated with them, no matter how trivial the relation is. People favour letters which appear in their own names, and things associated with their years of birth. 'Women were about 18% more likely to move to states with names resembling their first name than the should have based on chance' (Pelham, Mirenberg, and Jones 2002).

other whether Colin is blameworthy. I suspect the answer is complicated and somewhat measured. However, for the descriptive reflectivist, it seems that is the end of the story — Colin's decision simply fails to be one for which he is responsible and, presumably, he fails to be blameworthy. The normative reflectivist can say a little more. On that version of the view, we must ask 'How does closely does Colin's behaviour approximate the ideal of reflectively determined action?' But here we must address an ambiguity that has not yet been discussed. Is it enough for behaviour to approximate the reflective ideal that one act as the ideally reflective agent would act *irrespective of whether one comes to act that way a result of reflection*?

If we answer 'no', the view threatens to be unstable. Presumably it wouldn't be plausible to claim that it *doesn't matter* how one came to act in the way one did, so long as one's behaviour is similar enough to what the ideally-reflective agent would do. This is simply because there are obvious agency-impairing factors that could cause one to act that way, such as coercion or manipulation. But then we must ask 'What are the acceptable ways of coming to approximate the ideal beyond approximating it by reflection?' And this just seems to re-open the demand for a theory of responsible agency, one that needs to not be rooted in the notion of reflection. One could appeal to reasons-responsiveness here, and I think it would meet this demand well. We could say that one's behaviour is agential to the extent that it approximates the behaviour of the ideally reflectively determined agent, but that such behaviour must exhibit reasons-responsiveness. But then, naturally, we are led to wonder why we shouldn't just jettison the idea of approximating the action of an ideally reflective agent altogether, as it seems to be doing very little work, and just go in for a thoroughgoing reasons-responsiveness account.

If we answer 'yes', then there is less daylight between the normative reflectivist's position and the descriptive reflectivists's position than we should like. On this interpretation the normative version becomes a gradable verson (with threshold, perhaps) of the descriptive version. Reflection remains what matters for agency, but we have become willing to admit that it comes in degrees and that one needn't be perfect in order to be an agent. It would perhaps take some more caseology to show that this picture would or would not be extensionally adequate.[17] But I think, quite apart from that, that it is in the wrong spirit. Reflection may be one of the ways we exercise our agential capacities, and there are some reasons that we can only respond to effectively if we engage in reflection. But there are many cases of agency that don't seem to involve that sort of reflection at all, or cases where, insofar as reflection is involved, it seems to be of precisely the wrong sort (such as the Huck Finn case). Further, the picture of the mind that has only just begun to come into clear focus from the human sciences pushes us not only to re-evaluate whether our theories of agency are extensionally adequate, given what has been discovered. That picture should also push us to re-evaluate our ideal of human action, and to acknowledge that action and agency may be multifarious phenomena which, just like the mind, are not dominated by conscious

[17] I suspect this wouldn't work out so well for this version of the view precisely because of the sorts of examples furnished by the empirical literature.

deliberate reflection and our ideals of action and agency should reflect this acknowledgment.

What does the extended reasons-responsiveness account have to say about Colin? To say that the potentially agency-impairing factors are defeaters in Doris' sense doesn't settle the question of his responsibility according to such a theory. This is because how those factors are operating will be relevant for determining whether they are responsiveness-undermining. One advantage of the reasons-responsiveness account is that it allows us to distinguish different clusters of reasons with respect to which an agent may or may not be responsive. (This feature will help us in our discussion of addiction in Chapter 5.) For example, if Colin is subject to implicit egotism, he is presumably not responding (neither *de dicto* nor *de re*) to the reason "The name 'COLorado' is related to me...', nor to the reason "*This* name appears first on the list...' because those are not reasons to move to Colorado, nor does he take them to be. However, the ordering effects, or the effect of implicit egotism may be to allow Colin to become sensitive to facts which *are* reasons to move to Colorado. They may, for example, cause certain prominent and attractive Coloradan features to become more salient to Colin in his deliberation. On the assumption that those features of Colorado really do constitute reasons in favour of Colin moving there, it looks like we have a factor which is both a defeater in Doris' sense, and an enhancer of Colin's agency!

It if of course also true that the operation of such effects could serve to *diminish* Colin's responsiveness to reasons having to do with his daughter's well-being. Whether this should incline us to think that he is more blameworthy or less blameworthy depends on whether his failure to be so responsive is an example of indigent responsibility. So, assessing whether Colin is responsible is a complicated matter, on this view. We must take into account whether and to what extent the three potentially agency-impairing factors are indeed impairing (or perhaps enhancing!) Colin's agency. This involves assessing and tallying their effects on both his *de dicto* and his *de re* responsiveness to many different clusters of reasons. Even if we find that his reasons-responsive capacities have not been sufficiently exercised, we still won't have resolved the issue because it is still possible that Colin's behaviour is an example of indigent responsibility: he may have taken the decision too lightly, and without being sufficiently judicious. This may have been in part due to the operation of some agency-impairing mechanisms, but if those mechanisms are such that (and to some extent those mentioned are like this) their effects can be overridden with the exercise of careful reasoning, then we may find that Colin is blameworthy.

This is complicated picture, but I think it is as it should be. Reflectivism in all its forms seems to give answers too easily in cases such as this. The suitably extended reasons-responsiveness theory respects the complexity of cases like these (and many real-life cases are bound to be even more complex) where many factors are in play. And perhaps, above all else, that is the real lesson of the empirical results. There is a lot going on in the mind, and we have to be sensitive to it if we are going to be effective theorizers of morally responsible agency.

# Chapter 3

# Self-Deception

*You do it to yourself, you do,*
*and that's what really hurts*

RADIOHEAD, 'Just'

## 3.1 The Phenomenon

'Self-Deception' is a label that we give to a particular kind of motivated human irrationality. Although almost every element of the phenomenon is philosophically controversial — including its very possibility — if anything is a case of self-deception, this is:

> **Case:** D. is losing his hair. But if you ask him about it, he says, with seeming sincerity, 'I don't believe I'm going bald'. From a certain perspective, it's not altogether unreasonable that he would believe this because he doesn't encounter very much evidence pointing to his hair loss. Still, his friends harbour the suspicion that he *ought* to know better, that maybe on some level he *does* know better. 'Perhaps,' they think to themselves, 'he avoids mirrors or tells himself that the hair in the sink belongs to his wife. Maybe that's why he always wears a hat. After all, it would be a very painful thing to have to admit that one was going bald.'
>
> – *Adapted from* Davidson (2004a)

What is philosophically vexing about self-deception is that it seems to be something we that we encounter all the time — indeed something that we *do* all the time — but it is also notoriously difficult to describe in non-paradoxical terms. This difficulty is large enough that it has led some philosophers (Borge 2003) to deny the very possibility of self-deception. I do not think we should deny this phenomenon. Self-deception is both highly prevalent

and morally complex. Through its operation we subtly collude in the erosion of our own epistemic agency, and sometimes lead ourselves to ruin. But we also make use of it as a mechanism for coping with the mundane, and for moving ourselves where otherwise it might seem impossible. In order to vindicate the possibility of self-deception we must address its apparent paradoxicality. I wish to begin by situating the phenomenon of self-deception within the nexus of other related irrational phenomena. It is important to get clear on the precise sense in which self-deception is a species of motivated irrationality, and what makes that so. Then I propose to address the apparent surface-level paradoxes: how can someone intentionally get himself to believe what he knows to be false? There are three key notions at work in generating the apparent problem with this idea: belief, intentional action, and self. Correspondingly, there are three strategies for explaining the possibility of self-deception: we must revise what we mean, for the purposes of explaining self-deception, by 'belief', 'intention', or 'self'. I canvas the strategies of all three types as found in the literature, and find them wanting. Either, they fail to capture self-deception as a genuinely intentional phenomenon, or they fall back into the surface paradoxes. Ultimately, I want to reject all three, and to defend what I call Self-Deception as Omission. This view is motivated by trying to do what I claim that the three families of views canvassed cannot: (1) making sense of self-deception as something that we *do* and (2) drawing the lines of responsibility in the intuitively correct places. These features seem to be part of our ordinary conception of self-deception and are worth preserving.

## The Genus: Irrationality

What D. is doing in our case is clearly irrational. He is engaged in a process of belief formation which results in beliefs that are not consistent with evidence which it is not beyond his intellectual ability to obtain and appreciate. Generally speaking, irrational belief forming processes come in two varieties: those in which an affective or emotional factor is causally relevant to bringing about the belief, and those in which there is no such causally relevant factor. If we think that what D. is doing is an example of motivated irrationality, this already allows us to characterize the particular sort of irrationality that he is engaged in: on the assumption that the affective or emotional factor which is causally relevant to D.'s forming the belief that he is not going bald is *not* a good reason for believing that he is not going bald, then we can say that he allows something which is not a reason for belief to be its cause.

I say D. 'allows' this affective factor to be the cause of his belief because it will be the central thesis of this chapter that self-deception is something which, in some sense to be made more precise, an agent *intentionally* brings about (and I take it that allowings are intentional). Remarking first off that self-deception is a species of motivated irrationality takes us some of the way there already, since our affective responses, and how we manage, control, and respond to them as rational agents are things, on the face of it, for which we are responsible. I will have plenty more to say about self-deception and moral responsibility throughout this chapter and into the next.

Self-deception is most decidedly an example of motivated irrationality. This is not a trivial observation. The discovery of the difference between motivated and unmotivated irrationality marks real progress in our understanding of the failings of human reason since Freud. Freud appreciated the difference between failures of competence or brutely caused twinges on the one hand, and reasoned behaviour — of which he strongly believed irrationality to be a proper part — on the other. We can count it among his accomplishments that he succeeded in greatly increasing the scope of rationalizing explanations by shifting many phenomena from the former category to the latter. According to Freudian theory, otherwise seemingly inexplicable parapraxes, compulsions, exaggerated fears, or even dream episodes, are explained as the result of the principled operation of what Freud called 'wish' — or what we might prefer to call motivation or desire[1] — in the dynamic system of the psyche. They are, as we might say, brought into the space of reasons. And so much appears to be necessary if we are to understand those acts among them which are genuinely irrational. Irrationality is not, as Davidson repeatedly says, non-rationality. It is, 'a failure within the house of reason' (Davidson 2004b, 169).

However, we now know that not all examples of irrationality, and certainly not all examples of the phenomena Freud was interested in, have psychodynamic explanations. It is quite widely acknowledged that many types of psychopathology are not primarily due to the clandestine operation of affective states, but are much better explained by adverting to functional or organic causes. More to the point, it is now even widely believed that the 'softer' cases of irrationality do not have motivational explanations. There are many quotidian examples of these, which Al Mele calls, 'cold biases'. For example, consider what cognitive psychologists call the availability heuristic. When tasked with making judgments of probability or frequency, people are often led seriously astray by giving undue weight to things that are easier to recall, or are more salient. While it's true that if an event is more frequent, it will easier to recall, the converse does not hold. Nevertheless, the availability heuristic operates in accordance with just such a invalid principle. In a famous experiment, Amos Tversky and Daniel Kahneman (1973) asked subjects to judge the relative frequency of words beginning with the letter 'k' against the frequency of words with 'k' in the third position. Subjects consistently judge that words beginning with 'k' are more frequent, despite the fact that they occur only somewhere between one third and one half as often as words with 'k' in the third position. Tversky and Kahneman concluded that this was because when faced with the task, subjects set about recalling words in each of the two categories. Since it is much easier to recall words that begin with the letter 'k' (perhaps lexical search really does proceed, as it were, lexically) subjects concluded that those words were more frequent.

Another slightly more famous example of cold bias, also from Tversky and Kahneman, is as follows. Consider Linda. She is single, outspoken, and bright. As a student, she

[1]Some philosophers want to distinguish desires from mere wishes on the grounds that only the former, and not the latter, are subject to the seeming rational requirement that to will the ends one must will the means. I can wish to be the President of the United States, but if it turns out on interrogation that I am willing to take absolutely no further action in pursuit of this goal, one might come to wonder where I really desire it.

majored in philosophy and was deeply concerned with issues of social justice. She has also participated in anti-nuclear demonstrations in the past. Which is more probable? (1) Linda is a bank teller, or (2) Linda is a bank teller active in the feminist movement?

Despite the fact that it is a blatant instance of the conjunction fallacy to do so, a majority of subjects respond saying that it is more likely that Linda is a feminist bank teller. The explanation is that subjects use a resemblance heuristic to choose the option which more coherently resembles the stereotype associated with the description given of Linda. That is, they choose the option which seems to make for a more plausible story, rather than choosing the one demanded by the laws of probability.

One thing that characterizes these 'cold' biases and distinguishes them from self-deception, and perhaps from motivated irrationality in general, is that they appear to show no sensitivity to subject matter. For example, the following case, from Kahneman (2011, 7), parallels exactly the Linda case, but describes a person with different features and offers different occupational choices:

> An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has need for order and structure, and a passion for detail." Is Steve more likely to be a librarian or a farmer?

Even though there are 20 male farmers for each male librarian in the United States, respondents exhibit a strong bias towards thinking that Steve is a librarian. And just as in the Linda case, the explanation is that respondents use a resemblance heuristic to match the description they have been given to a stereotype, in this case an occupational stereotype. It appears that the operation of this biased mechanism is indifferent to the nature of the descriptions given of Linda or Steve (and the concomitant stereotypes those descriptions conjure or are associated with) and to the particular nature of the occupational choices they were given. It would be implausible to think that the explanation for the biases in these cases had something to do with a heretofore unknown but very prevalent set of positive or negative affective associations with bank-tellers, feminists, farmers, or librarians. The tendency to substitute one easier task ('How much does this description resemble the farmer/bank-teller stereotype?') for a more difficult task (calculating probability) is perfectly general — at least as far as subject matter is concerned. The same, of course, is true of the availability heuristic. People tend to judge as more frequent anything which is more salient or easier to recall: airline travel is deemed more dangerous when there has been a recently publicized aviation accident; smoking is deemed less likely to cause lung cancer by people who can recall a long-lived relative who smoked; the names of famous people seem more common; words starting with 'r' or 'k' are deemed more frequent that those with 'r' in the second position, or 'k' in the third.

## Differences Between Cold Bias and Motivated Irrationality

What is characteristic of motivated irrationality is that an affective dimension somehow contributes to that which makes it irrational. *Akrasia*, or acting against one's own better judgment, is clearly like this: without the motivation to act in the suboptimal way, the *akrates* would have no struggle to contend with. The motivation is allowed to cause the action even though it is not a decisive reason for the action. Wishful thinking is also clearly like this. In wishful thinking, a subject allows a desire or wish that something be the case to be the cause of a belief that it is indeed the case. Again, something which is not a (decisive) reason is allowed to be a cause. Self-deception can be seen as a stronger version of wishful thinking. What D. does is not only to allow his desires that something be the case to cause him to believe that it is the case in the absence of evidence, but also to cause him to believe it *in the face of evidence to the contrary*. As an even stronger version of wishful thinking, self-deception is clearly an example of motivated irrationality.

In contrast with the various cold biases, it is characteristic of motivated irrationality that its operation be directed towards a very specific subject matter: that with which the subject is emotionally entangled.[2] Wishful thinking, for example, must be wishful thinking that something is that case, something which the subject *wishes* to be the case. It would scarcely make sense to think that there could be a *topic insensitive* mechanism for producing beliefs which a subject has a vested emotional interest in having. Nor would it be plausible to suppose, for example, that in *akrasia* the akratic subject acts as he judges he ought not to because of the operation of some generally biased action-producing mechanism. If it is true that there is a general psychological propensity in human beings towards eating the cookies, even when we judge we ought not, it is only because of the (perhaps quite prevalent) desire to eat cookies which is telically keyed to *cookies*. So, it is not in the least bit puzzling that someone in a situation rife with cookie affordances and saddled with very strong desires to eat cookies may find himself behaving irrationally in a way that involves cookies and not in a way that involves baseball cards.

A further difference between cold bias and motivated irrationality can be seen when we examine what happens when the irrationality in question is discovered by the subject. People who are engaged in biased heuristic reasoning are typically not aware that they are doing so. In fact, by hypothesis, these heuristics operate below the level of conscious awareness. They are elements (as readers will recall from the previous chapter) of System 1. System 1 is very good at providing swift and effortless navigation through a host of everyday situations and at offering up intuitive answers to simple cognitive problems. These answers are, as the research program has extensively indicated, often systematically biased. But avoiding the pitfalls of these biased mechanisms often takes nothing more that the mobilization of so-called System 2. The exercise of System 2 is conscious, deliberate, slow, and effortful but can be used to successfully detect, and then to decisively overthrow, the biased workings of System 1. All a subject needs to do, in the examples discussed above, is to catch herself

[2] I will have more to say about how I understand what it means for a subject to be 'emotionally entangled' with some proposition in the following chapter.

substituting the easier task for the more difficult one: if she put in the effort to draw the Venn diagrams depicting the logical relation between the set of bank-tellers and the set of feminist bank-tellers, or if she carries out a more systematic counting procedure to keep track of the relative frequency of 'k'-initial words and words featuring 'k' in the third position, she will come to clearly see the correct answer. Often it is enough for the subject to have the correct answer pointed out to her — that may be enough to encourage her to engage System 2 to recognize the nature of the error. (Understanding and recognizing the operation of biased mechanisms is, of course, a System 2 activity.)

This differs quite sharply from most examples of motivated irrationality. Pointing out to someone that they are engaged in wishful thinking, for example, often does nothing to extinguish it. And this we should find unsurprising. As an example of motivated irrationality, wishful thinking has a desire as part of its cause, and becoming aware of the presence of such a motivation needn't do anything to eliminate it or its effect on the subject's cognitive states. Likewise, in *akrasia* the subject already knows that what he is doing is contrary to his better judgment — pointing it out to him does nothing to change his situation, except perhaps to register disapproval.

What this seems to show is that the source of motivation which is operative in wishful thinking and in *akrasia* is in some sense independent of rational judgement. This is why rational unmasking does nothing to extinguish the operative motivation. In self-deception the operative motivation is also independent of rational judgement, but since self-deception doesn't work unless it remains clandestine, we can't come to see this by asking whether the irrationality survives unmasking. If it becomes plain to a subject that she is deceiving herself, her self-deception will fail, but not because the *source* of her irrationality has been eliminated. Often a failed attempt at self-deception will bring about a pattern of *thinking wishfully*,[3] where the agent may even rue the failure of the self-deceptive attempt and pine — now explicitly — for the truth of what she wishes for. This would not be expected if the motivation did not persist.

So far I have attempted to identify a couple of key features that distinguish self-deception from other kinds of irrationality.

1. It's not mere *error*. If I fail to see that something about my behaviour means that some discomforting thing is true about me due to a failure of competence I'm not self-deceived, I'm merely self-ignorant.

2. It is highly *selective*. The various 'cold-biases' operate regardless of subject matter. But self-deception involves a mechanism which is highly selective: D. ignores (or is insensitive to) only things that tend to indicate that he is balding.

3. The source of motivation is independent from — and must in some sense be concealed from — rational judgement.

[3]As distinct from wishful thinking. The way I have glossed wishful thinking here — believing in spite of lack of evidence — it too would be ruled out by the unmasking of a self-deceptive attempt.

The first two of these features place us at once in the sphere of potentially reasoned irrational actions, as opposed to brute twinges, and distinguishes those among them which are motivated from those which are not.[4] They are also related: Self-deception is motivated, and it is targeted, but the motivation *specifies* the target; self-deception is driven by a motivation which is telically keyed to a specific propositional object. The network of rational relations surrounding that propositional object comprises the set of things to which the self-deceived agent is insensitive.

The third feature is, I have claimed, a feature shared by species of motivated irrationality, with the added proviso that there is additional layer of complexity produced by acknowledging the fact that the operation of self-deception must be clandestine. This feature of self-deception is obviously crucial, and will arise again several times as we proceed.

## 3.2 The Surface Paradox(es)

That self-deception is to be located in the genus of motivated irrationality is not particularly controversial amongst philosophers and psychologists today. But almost every other possible dimension of the analysis of this phenomenon is fraught with disagreement. And it is not surprising that this should be so. Our ordinary experiences with self-deception, in ourselves, and in others, bring us face-to-face with many complex and vexing issues concerning agency, and mentality more generally. Two extremely important philosophical concepts which are directly interrogated by the phenomenon of self-deception are intentional action, and belief. A brief remark about each of these notions before moving on: First, intentional actions, roughly speaking, are instances of goal-directed agency, and are to be distinguished from mere happenings, twinges, twitches, and ticks and the like. Second, beliefs are states that purport to represent the world, and form the basis for our reasoned speech and action. Because the *telos* of belief is to represent the way the world actually is, it is generally agreed that belief is not (except perhaps in cases of positive thinking and related phenomena)[5] under the control of the will. Nevertheless, when we (at least naïvely) remark that D. has deceived himself, we are tempted to say *both* that he believes that he is losing his hair and that he is not losing his hair *and* that he believes the second thing for a motivationally biased reason.

[4]For the moment I wish to dodge the question of whether the operation of unconscious biases counts as reasoned action or not. This is part of the reason why I began the discussion with a mention of Freud, whose attempt to explain all irrationality as a kind of reasoned action was at the very least very gallant. The point at present is just that we are talking about irrationality, not non-rationality.

[5]It is perhaps worth noting that the problem with producing beliefs at will is not that willing, generally, cannot produce beliefs, it is that normally the willing is not a good reason to hold the belief. But this is compatible with there being cases where the mere holding of the belief actually does constitute a (defeasible) reason to think it true. So if it really is true that believing that I can jump the chasm (Johnston 1988) increases the likelihood that I can in fact jump the chasm, then there my be no problem with my acquiring this belief 'at will'.

On a very natural description of self-deception, belief and intentional action are related as follows: the agent believes $p$, and intentionally brings it about that she believe its opposite. This very natural description comes about by thinking of self-deception as a very special case of other-deception, where the deceiver and the deceived happen to be identical. This is an attractive starting point at least in part because a *prima facie* satisfactory analysis of other-deception doesn't seem too hard to come by: A deceived B that $p$ just in case A knows that $p$ is false, but has intentionally caused B to believe that $p$. Being confident in this much isn't nothing. But it also seems to capture quite nicely the idea that the self-deceiver is responsible for being the deceiver (and perhaps also that he is responsible for being duped). In ordinary other-deception we know very clearly how to draw the contours of blame, and it is an important virtue of any account of self-deception that it draw these contours where we think they belong.

So, we might be tempted by the following view of self-deception:

**The naïve view of self-deception:** A is self-deceived that $p$ just in case A believes that $p$ and A has acted intentionally so as to cause A to believe not-$p$

But we can see already how if this is our starting point, we will quickly run into difficulties. At the interpersonal level, the role played by intentional vs. non-intentional action in cases of what we might call generic misleading is quite clear: The difference between *merely* misleading you into believing $p$ and my positively deceiving you into believing $p$ depends entirely on whether I *intended* to cause you to believe $p$.[6] And of course, my deception of you is at best incomplete if I have as yet failed to cause you to *believe* the thing in question. But how, given the kinds of things that we think belief and intentional action are, could this be the right understanding of the phenomenon? If we start with other-deception, although we have started on fairly secure conceptual ground, we have proceeded, with the 'mere' added stipulation that the deceiver and the deceived be identical, to describe something that scarcely seems possible. How can I *intentionally* get *myself* to *believe* something that I also believe to be false?

First, there seems to be something puzzling about the *state* that D. is in when he is self-deceived. One might have been tempted to describe it this way: On the one hand it seems like he believes that he is going bald, and on the other hand, it seems like he does not believe this. But this would make the difficultly all but insurmountable — it would be to say that D. believes that $p$ *and* that it is *not* the case that he believes that $p$. If anything is a contradiction, this is. So if there is a *static* problem, this can't be the right way to describe it. We might quite plausibly think it better to describe the puzzling state that D.

[6] There is a complication here: in order for either kind of 'generic' misleading to be misleading at all you need to have been caused to believe something *false* as a result of my actions, but the thing in question needn't be $p$. It seems that I can deceive (or mislead you) into believing $p$ even though $p$ is true, if, for example, I intentionally (or non-intentionally) cause you to believe *something else*, $q$, which is false, but which you take to be a sound basis for believing $p$. I don't believe that this complication bears on anything that comes below.

is in as the one where he simultaneously believes $p$ and believes not-$p$. But is this really all that puzzling? For this to be possible, all that would have to be the case is that we sometimes have beliefs that are inferentially insulated from one another. And this should not strike us as the least bit implausible. The mere possibility that the body of my beliefs is not maximally logically coherent forces this on us. Even if (wildly contrary to fact) I had consciously inspected each one of my beliefs for accordance with evidence and for coherence with my other beliefs, it is still highly unlikely that no such incoherence or lack of support would escape my notice. In any case, it is wildly implausible to suppose that I have ever done any such thing. I have many beliefs that I am not aware that I have, and it is almost certain that they don't constitute a consistent set.

The real problem with the naïve view is not that it says self-deceivers must in some impossible or particularly puzzling state. Rather, the problem is a *dynamical* one. How can I, as a more or less unified agent intentionally come to be in a state where I believe not-$p$ *from* a state where I believe $p$ (where my belief that $p$ persists)? The trouble is not with the idea of *belief* or *intentional action* per se, but rather with the particular way in which they are said to interact in a single more-or-less psychologically unified agent. There are three moving parts to the problem: belief, intentional action, and psychological unity. *If* the self-deceived agent could somehow pull off an intentional act of getting himself to believe what he also believes to be false, wouldn't this undermine his psychological unity? This seems like a process which being unified would suffice to foil. *If* what the self-deceived sufficiently unified agent manages to bring about in himself is a genuine *belief* that not-$p$, wouldn't he have to have done it non-intentionally? After all, belief isn't (normally) under the control of the will. *Mutatatis mutandis* if the sufficiently unified self-deceived agent *really intends* to deceive himself, how can the deception involve the bringing about of a genuine belief? Beliefs aren't (normally) under the control of the will.

Much of the philosophical literature on self-deception is organized around trying to find a solution to the dynamical problem, and the available views can be sorted according to how they depart from the naïve view. Understanding the problem in this way helps us to see the three main strategies that various philosophers have embraced to solve the dynamical problem: (1) deny that one of the beliefs in the contradictory pair rises to the status of full belief; (2) deny that the 'self' involved in self-deception is unified; and (3) deny that the act of self-deceiving really counts as fully intentional.

I choose to deal with the dynamical problem in a different way. Rather than trying to tweak the content of the self-deceptive intention, or the precise nature of the representational state that the self-deceiver is in, I propose to reanalyze the structure of the self-deceptive process. Oddly, the theorists who have tried to take seriously the dynamical problem that the naïve account seems to face have not thought to alter the form of agency that the naïve account imputes to self-deceivers — that is, roughly, intentional action of some kind. According to my view self-deception does not involve intentional action, but rather intentional *failure* to act, intentional omission.

My account is both constrained and motivated by capturing another very important aspect of self-deception. Self-deception is more than just willful belief formation. It is also

a type of self-induced ignorance, and there is thus a very strong *prima facie* presumption that self-deceivers are *responsible* for their self-deception. This is, in part, what makes self-deception a phenomenon of concern for moral psychology: can we make good sense of self-deception in a way that justifies the attitudes that we hold towards those who perpetrate it on themselves? Of course, doing so will necessarily involve making sure we have an account of the phenomenon which makes it psychologically possible, but that is not the only desideratum. Ordinary moral thinking seems to hold that self-deceivers are responsible and that blame is *prima facie* appropriate. If there is one important thing that the naïve view gets right, it is this: there is a clear node of agency, to wit, an intentional action, on which we can hang our judgements of blameworthiness.

Solving the dynamical problem thus ought to go hand in hand with preserving the idea that self-deceivers are responsible. However, as I will try to show as we go along, the competing goals of responding to the dynamical problem and holding our responsibility judgements intact are in tension. The way to resolve the tension, and to satisfy both desiderata, is to find the locus of agency in self-deception not in an intentional action which is identical with the act of deceiving oneself, but instead in an intentional omission, which only *partially* constitutes the process of self-deception as a whole. Simply omitting to do something isn't going to be enough to get an agent into the self-deceived state. *What* the agent omits to do is to intervene in a biased, unconscious, subpersonal belief forming process. That process gets the agent into the state of having the self-deceptive belief, but what the agent is responsible for, the locus of agency in the phenomenon, is omitting to intervene in, or to overthrow the result of, that process.

First, I will outline my view in more detail and show how it both avoids the dynamical problem and holds intact our responsibility judgements. Then, I will turn my attention to critical discussion of the other main views on offer in the literature.

## 3.3 Self-Deception as Omission

My view of self-deception is inspired by now-familiar dual-process psychology. According to my account, which I call Self-Deception as Omission, the episode of intentional agency for which the self-deceiver is responsible and the process of belief formation come apart. According to this view, the agent is not responsible for the formation of the self-deceptive belief itself, which occurs unconsciously (as the result of a certain type of 'System 1' process),[7] but rather she is responsible for acquiescing in that belief once it has been formed, where the acquiescence itself is a voluntary ('System 2') *omission*. Let me elaborate.

We are familiar enough with System 1 and System 2 from Chapter 2. But to bring it vividly back to mind, consider the following example from Kahneman. Again, his instructions are to 'listen to your intuition' (2011, 44):

[7]Or may occur as a result of such a process. I consider a weakening of my view which doesn't require that the belief have this source in the following chapter.

A bat and a ball cost $1.10.
The bat costs one dollar more than the ball.
How much does the ball cost?

The intuitive — incorrect — answer that System 1 offers up is $0.10. And it seems to do so more or less unbidden, once the specification of task has been grasped. Whether one chooses to go on and perform the calculation and ultimately arrive at the correct answer seems to be an independent matter. What happens is that *first* System 1 issues its verdict, and then, if one chooses, one can go on to perform the calculation and then *override* the initial judgement offered up by System 1. What is most important for our purposes is that System 1 has caused the agent to have a *belief*. Kahneman tells us that among the automatic activities performed by System 1, we find

1. Detecting that one object is more distant than another
2. Orienting the source of a sudden sound
3. Completing the phrase "bread and..."
4. Making a "disgust face" when shown a horrible picture
5. Detecting hostility in a voice
6. Answering 2 + 2 = ?
7. Reading words on large billboards
8. Driving a car on an empty road
9. Finding a strong move in chess (if you're a chess master)
10. Understanding simple sentences
11. Recognizing that a "meek and tidy soul with a passion for detail" resembles an occupational stereotype

Most of the activities seem to involve belief-formation. With the exception of (4), all of these activities seem to involve propositional attitudes. (1) and (2), and perhaps (5) seem to involve beliefs which are very much akin to perceptual beliefs: that the sound came from over there; that this object is closer than that object; that the person on the phone is hostile. (7) and (10) seem straightforwardly to involve the belief that the things being read say what they in fact say; (6) that the answer is 4; (9) that such-and-such is a strong move etc.

So it seems we can accurately characterize the operations of System 1 as what we might call *quasi-intentional*[8] in the following sense: System 1 aims at something — at least some of the time at the production of beliefs — but its operations are unconscious and are not themselves practically syllogisable. That is, its operations are not effectively modeled as actual or potential episodes of *reasoning* from propositional states of the system e.g., beliefs or desires — the system itself does not have these states, although it has access to some of those that are properly said to belong to the whole agent, and can produce them. What the results from the empirical literature appear to show is that System 1 aims not just at the production of *true* beliefs, but at the rough-and-ready production of for-the-most-part true beliefs. It is widely believed that the evolutionary value of a fast and highly adaptive parallel system outweighs the disvalue of sometimes getting it wrong when it comes to, e.g., abstract problems about probability. When the goal is to believe in accordance with the norms of probability theory one must be more judicious in listening to what is offered up by intuition. Indeed, what happens in cases like the 'Linda' example is that subjects simply fail to perform rational self-checking before offering their answer and end up expressing a belief whose cause (proximate: System 1 operation; distal: evolutionary pressures selecting for 'true-enough' over 'true') is not a reason in the standard normative sense for a belief about probability. What makes this a failure of rationality is that the subject is not sufficiently judicious on the uptake.

What self-deception seems to involve is the operation of a quasi-intentional mechanism for the formation of fairly specific sorts of beliefs: those with which the subject is in some way or another emotionally entangled. Typically, System 1-type biases are 'cold' in the sense that they don't seem to involve an affective component and are subject matter general. But it is quite plausible to think that there are quasi-intentional psychological mechanisms for regulating our emotional lives by regulating beliefs about, e.g., ourselves, our relations to things that we value, etc. We could call these mechanisms 'doxastic affect-regulating mechanisms' (DARMs).[9] These mechanisms are analogous to System-1-type belief forming mechanisms precisely in the sense that they too are quasi-intentional. They aim at something without being conscious or practically syllogisable. The difference is that what the DARMs aim at is affect regulation, instead of, or perhaps in addition to, an evolutionarily balanced combination of efficiency and approximate truth, and they do so via belief modulation. Of course, regulating belief is not the only way to regulate affect, but many of our feelings, even our deepest feelings, reflect what we believe about the things we care about, and where those things are concerned — as they often are in cases of self-deception — belief modulation is a clever (but subpsersonal) strategy for emotional or affective regulation.

How might one come to be self-deceived via the operation of a DARM? The quasi-

[8] A note of clarification about 'intentional' and its cognates: I do not mean Brentano-style 'aboutness', but rather that features of certain actions which they possess in in virtue of being voluntarily undertaken for the sake of accomplishing some end. 'Quasi–intentionality' is thus not a form a quasi-aboutness, but rather a logically weaker kind of goal-directedness.

[9] Mild apologies are in order for this, but there are affect-regulating mechanisms which don't work via the production of beliefs, e.g., drinking one's morning coffee, or having a disposition of temperance.

intentional operations of the mechanism produce in the agent the comforting belief, for example, that he is not going bald by causing selective attending to evidence, biased evaluation, or by exploiting whatever other mechanisms are available. Consider the following re-description of D.:

> **D.*** D. is pretty clearly going bald, though if you ask him what he thinks of the matter he will deny it, with seeming sincerity. Without being aware of it, D. was caused to believe that he is not going bald by processes operating in him, processes which made him, perhaps: avoid situations where he might encounter evidence of his hair loss; prefer the way he looks from certain angles in the mirror; not think twice about the hair in the sink etc. Although he himself can't be quite sure how he got his belief, it is nevertheless doing something for him. He feels much better about himself than he would if he believed he was losing his hair, and for that reason, on a conscious, personal level, he treats the matter as closed.

This is the sense in which self-deception, on this view, is quasi-intentional: it comes about as the result of the operation of a quasi-intentional process, a process which aims at something but is subpersonal. This is how we solve the dynamical problem: the belief can come about in the agent even though he never intends to get the belief to come about in him. And we are able to get this result without doing violence to the intuition that self-deception is something that the agent had perpetrated on himself. Once the belief is present in the agent, the quasi-intended affective results can begin to be felt. If evidence against the belief is available, in the sense that a reasonably judicious look into the matter would reveal it, but the agent is caused to forswear further investigation by a desire to remain in the affectively more palatable state when he otherwise would proceed, we have a case of self-deception. What is *irrational* about the agent's accepting the belief is that he acts as though the cause of his belief and its effects — the operation of the DARM and its boost to affect — provide good reasons to continue to hold it. What he *does*, the thing for which he is responsible, is to acquiesce in the affectively more palatable state that the affect-regulating mechanism has offered up.[10] According my view, then, one becomes self-deceived via default action for which one bears indigent responsibility (§2.4). Continuing to believe as the DARM has caused one to believe is something that requires the exercise of System 2 in order to prevent, but it is also something one ought to prevent. That is, when the evidence against what one believes is strong and readily available, one is responsible for failing to do what is necessary in order to recognize and appreciate it.

[10]It may have occurred to the observant reader to ask at this point why my view is not a version of views according to which we give up the unity of the agent to solve the dynamical problem — views which I will later call 'fracturing' views. It has many affinities with such views and may, technically speaking, be in that class. However, the fracturing that my view appeals to is not *ad hoc* and is instead motivated by more systematic considerations, such as the distinction between type-1 and type-2 processes (especially as understood on the natural kinds hypothesis from §2.3). See below ('fracturing') for discussion of the affinities, as well as my critique of a popular competing token view of this type.

A word about the sense in which the evidence against the belief must be 'available': I mean this, perhaps obviously, to be an 'externalist' notion of availability. It may be false that the agent 'has' the evidence, in the sense of already having apprehended and appreciated it. But it may nevertheless be true that the evidence is there, in the world, ripe for discovery. I consider this externalist notion of evidential availability to be more appropriate for an investigation of the ethics of self-deceptive belief simply because it seems to me that an internalist cannot correctly identify what is irrational about self-deception. Self-deception is, among other things, to believe, for motivationally biased reasons, *in the face of evidence to the contrary*. But, it seems that an internalist about epistemic reasons would have at least a *prima facie* difficulty making sense of how the self-deceiver has reasons to not believe as she does when precisely what is distinctive about her is that she has somehow managed to make those very reasons obscure to herself.[11] It is an interesting consequence of thinking of self-deception in externalist terms that the difference between self-deception and wishful thinking becomes a matter of externally determined degree. Intuitively, where self-deception is believing in the face of evidence to the contrary, wishful thinking requires only belief in spite of a lack of evidence in support. If this account is correct, however, the quality and abundance of the evidence available to the agent in the externalist sense can vary independently of whether the subpersonal mechanisms and the corresponding voluntary omission that are distinctive of self-deception have been employed or perpetrated. Two agents who are the same 'in the head' can be such that one is a self-deceiver and the other is merely a wishful thinker if the former is in a situation where the available evidence speaks strongly against the belief she has and the latter is in a situation where the evidence is weak or equivocal. I consider this a feature rather than a bug.[12]

[11]There is a worry here. If self-deception requires that the agent believe something which is externally defeated, does this mean that it is not possible to have self-deceptive beliefs which are true, and are robustly supported in the external sense, but where the agent believes 'accidentally' or in a way which is not sensitive to the evidence which is there? (This is an interesting inversion of a worry that typically dogs *internalist* theories of justification generally.) My answer is that I accept that this may not technically count as a case of self-deception on my view. I accept this cost in light of two considerations. The first is the point just mentioned in the main text. If the notion of justification at issue is 'by the lights of what the agent already believes, and what evidence the agent has already appreciated', I don't see how we are going to get out of the dynamical problem. So it might just turn out that the self-deception-by-internalist-lights imagined by the objection just isn't possible. The second consideration softens the blow of accepting that such self-deception wouldn't count as 'true' self-deception on my view, if it turned out to be possible. If what the agent believes really is true, and really is well supported, externally speaking, then I am inclined to think that the moral risks and possible costs of having such a belief are simply going to be lower and fewer when compared with cases of 'true' self-deception. We should then expect that this will be reflected by less reproachful reactive attitudes directed at the agent. Insofar as I have been stressing that self-deception meets with a distinctive kind of moral response, the fact that our responses are different for these 'shmelf-deceived' agents makes me think it would not be onerous to accept a slightly different classification for them.

[12]It may occur to some to run a *modus tollens* against my *modus ponens* here. The worry could be: there is some more significant difference between self-deception and wishful thinking than just the respective degrees of external justification. My reply to this is that the relation between wishful thinking and self-deception is to a certain degree up for grabs — although they definitely seem to be related. The fact that wishful thinking seems generally to invite lower-grade censure than self-deception suggests to me that they

So, on this view, there is no single act which is an act of intentional belief formation against a belief that one already holds. One may of course hold the negation of the self-deceived belief, but I don't think there is anything particularly puzzling about being in *that* state.[13] What was puzzling, rather, was how one could manage to get into it intentionally. The answer is that one doesn't get into it intentionally, but only quasi-intentionally. The nascently self-deceptive belief is produced by a mechanism telically keyed to certain psychological outcomes whose operation is below the level of awareness and is not properly understood as fully reasons-responsive. Nevertheless, the agent commits an epistemic faux pas of a distinctive kind if the affective quality of that belief is the cause of the agent's continuing to hold it despite the fact that it is not a good reason to do so. And it is this acquiescing in the belief that the self-deceived agent is responsible for. Decomposing self-deception into this complex two-part phenomenon solves the dynamical problem without failing to capture the way in which self-deceivers are responsible for being self-deceived.

Because continuing to believe the affectively more pleasant belief that already has is what will happen unless the subject performs a System-2-type intervention, self-deception, on my view, is a kind of default action. And this is what allows my view to capture the responsibility facts that I have been emphasizing it is important for an account of self-deception to get right. But, it is worth making clear that, according to my view, the thing for which the self-deceived agent is responsible is a voluntary omission. Now, not all omissions are straightforwardly things for which an agent is fully responsible. In typical cases where robust responsibility attaches to an omission, the agent is aware of the consequences of forswearing to act, as well as the alternatives — actions — and *their* consequences (within reason), yet chooses to refrain from performing any of them. Suppose I am minding my own business on a park bench as a trolley whizzes past. Suppose further that at that very moment a strong breeze parts the trees in front of me revealing both a helpless bystander trapped on the track and a lever with a sign reading 'Flip switch to divert trolley'. It seems to me that I would be at least partially responsible for the bystander's death if I didn't at least *try* to get to the lever in time. Contrast this with the case where I am aware of neither the bystander nor the lever. I am simply sitting reading a lovely article by Judy Thomson and minding my own business. Assuming my ignorance is itself excusable (nothing about the situation presented itself as abnormal to me, we may suppose), it is that ignorance which relieves me of responsibility

---

are related as phenomena on a moral continuum. Indeed, as I will say more about below, the degree to which an agent is blameworthy may turn out to involve a significant degree of moral luck and thus might be due to facts which are largely outside of the agent's control (because they are downstream effects etc.). This suggests that the difference between self-deception and wishful thinking *does* turn on matters 'outside the head.' The fact that we generally find self-deception more worthy of serious censure is that we generally think that believing in face of evidence to the contrary is a riskier proposition than believing in the absence of evidence.

[13]Indeed, there may be good reason to suppose that in most cases of self-deception, the self-deceiver *will* have such contradictory beliefs. For example, the DARMs may need to have access to the affectively disturbing belief (at least at the level of suspicion) to 'know' that belief in its negation ought to be brought about. But this doesn't seem implausible. As I noted above, the mere possibility that the body of my beliefs is not maximally logically coherent forces this on us anyway.

for failing to act to save the bystander.

Thus, one might worry whether any of the conditions that typically attend cases of robust responsibility for omission are present when an agent is self-deceived, on my view. It was precisely in response to the dynamical problem attending the naïve view that we were motivated to find a view where the agent *isn't* aware of what she is doing as an act of deceiving herself. Doesn't this undermine her responsibility?

I grant that in typical cases of omission certain awareness conditions must be satisfied by the agent. The agent cannot be held responsible for failing to act on an alternative she was excusably ignorant of. This seem to follow from some version of 'ought-implies-can': the agent's knowledge of the nature and availability of the alternatives is a condition on her choosing to act on any of them. But the requirement that the self-deceived agent violates is not one that applies to a deliberative situation *given* her epistemic state. Rather, it is a norm that applies to her epistemic state itself. She is not excusably ignorant of what she is ignorant of: she has acquiesced in the more comfortable, but externally defeated, belief. Exercising epistemic agency in accordance with the norms that govern it is an effortful process, and the results can be disappointing, or worse, but this does not excuse the agent from doing it. Thus, a failure to do so motivated by a desire to avoid effort, to avoid possible disappointment (or worse), and to remain in the comfortable ruddy glow of one's positive self-image (as the case may be), is a failure for which the agent is blameworthy. Ignorance may be enough to change how one is required to act in a given situation, but it cannot itself excuse the agent from the epistemic norms which determine whether that ignorance is itself blameworthy or not.

This is not to suggest that there is a hard and fast distinction between moral norms and epistemic norms. It may seem as though I am suggesting that there is a set of moral requirements that the agent gets out of because of ignorance, only to find himself subject to another set of non-moral epistemic requirements for which ignorance is no excuse. This is true, but suggests that the distinction between the two sets of norms is cleaner than I think it is. Rather, I think it behoves us to see that the violation of an epistemic norm can have attendant consequences which are of genuine moral significance. For example, one can bring about genuine harm to oneself or others by failing to do one's epistemic duty. In cases such as this, it seems clear to me that the agent can be blamed for the resulting harm, *and* the ignorance which was causally efficacious in bringing about the harm. This is not double counting. The agent is generically responsible for any action or omission properly attributable to her. Should that action or omission have foreseeable consequences which are both morally significant and negatively valenced, the agent is blameworthy for both the consequences and their cause *but* the blameworthiness for the cause is, as it were, inherited from the badness of the consequences. That an action or omission is attributable to an agent is thus a form of liability, and the epistemically irresponsible agent takes a moral risk in proportion to the potential seriousness of the effects of causes which may be, to a significant extent, beyond his direct control. The world is a risky place, and one must be prepared to bear the censure that properly accrues to one as a result of one's epistemic failures. Notice, however, that if there were a positive correlate of liability, it would not

apply in a symmetrical way. An agent is not praiseworthy if, as a result of some epistemic failure of hers, some unforeseen positive consequence results. Many theorists have noticed the various asymmetries between praise and blame, and here we see one of those asymmetries in a big way, ramified by the effects of moral luck.

The affinities with the (in)famous Cliffordian position will likely not have gone unnoticed (Clifford 1999). Clifford famously claimed that '[i]t is wrong always, everywhere, and for anyone to believe anything on insufficient evidence.' I take it that Clifford's position, like mine, relies on the truth of some version of a principle which allows moral responsibility to be transferred up causal chains, even when the first link of the chain is a distinctively epistemic failure. I do not, in fact, agree with Clifford that every instance of epistemic indigence is wrong (and hence typically blameworthy), but that is not because of a disagreement over the details of such a principle, but rather a disagreement about whether there is *bound* to crop up, somewhere down the line, some harm which traces back to the agent's epistemic failure. Sometimes, it seems, such agents can get lucky. It's just that when they are unlucky, the mere attributability of the epistemic misstep to the agent becomes genuine moral blameworthiness.[14]

## 3.4 Competing Views

I now wish to turn to critical discussion of some of the competing views of self-deception in the literature.

### Pretense

Another family of views about self-deception proposes to avoid the surface level paradoxicality by characterizing the attitude that the self-deceiver ends up with not as belief, but as something else which plays a similar role to belief. I will focus on Stephen Darwall's version of the view (Darwall 1988), but Tamar Gendler has also proposed a similar view (Gendler 2007) and Jason D'Cruz has a revision of Gendler's view (D'Cruz, In preparation). According to the self-deception-as-pretense view, when one is self-deceived about $p$ one need not believe it (although one does typically believe its negation). Rather, one is engaged is an elaborate *pretense* according to which $p$ is the case. One acts *as if* $p$ were true. But unlike in ordinary pretense, where one also believes that one is engaged in pretense, when one is self-deceived, one is also engaged in a *second-order* pretense about one's first-order pretense: one behaves *as if* one is not merely behaving as if $p$.

A number of interesting questions arise about this kind of account. Darwall says explicitly that his aim is to 'escape the paradox of construing the self-deception as simply an internalized version of other-deception' (Darwall 1988, 415). The problem, as Darwall sees it, is that such an account would commit us to the view that the self-deceived person 'literally

[14]The difference between these two kinds of responsibility will be central the discussion in Chapter 4.

believes what he knows to be false'. So it seems Darwall is motivated to avoid the dynamical — and perhaps even the less serious static — problem as well.

Darwall invokes a contrast between what one *thinks*, and what one believes. According to Darwall, when one acts against one's better judgement, for example, one needn't abandon one's belief about what would be best all things considered. Rather, in Aristotelean manner, he claims that what happens is that one is distracted or led off course by *thinking* about all the tempting aspects of doing otherwise. Similarly, Darwall claims that the self-deceived agent needn't literally believe the thing that he is supposedly self-deceived about. Perhaps he just *thinks* various thoughts that amount to a kind of elaborate pretense to the effect that the thing otherwise believed to be false is the case. But this wouldn't yet be enough. In ordinary cases of pretense, we *know* that we are pretending. So, in order for the pretense to have the desired psychological results (preservation of self-image, successfully avoiding facing up to painful realizations etc.) it seems that the nature of the pretense itself has to be concealed. As Darwall puts it '[This is] not simply the first-order pretense involved in fantasy, but also the second-order pretense that...pretensions are real. When the self-deceiver plays the role of fool to himself, he must also pretend that he is not playing that role' (Darwall 1988, 414–415).

This kind of account displays a sensitivity to the potentially distorting effects of motivation and affect and makes an interesting appeal to the idea of fantasy which resonates with a lot of our ordinary experience of self-deception. But basing an account of self-deception on pretense or fantasy seems, in the end, to face a version of the dynamical problem. I take it that the reason that the second-order deception is necessary is to conceal from the self-deceiver the fact that he is engaged in pretense. But now we have to face squarely the question of how someone could ever manage to get himself into *that* state in the first place. And further, the purpose of engaging in the pretense must not be simply to sharpen theatrical skills or for merry diversion. Plausibly the purpose is something self-directed and psychological such as the preservation of self-image, or to avoid facing up to some painful facts. But this is the sort of thing that just can't be achieved by pretense if one knows that one is pretending. So the purpose of engaging in the pretense, whatever it is precisely, will often be something which cannot be achieved unless it is hidden from the agent. Darwall himself is acknowledging this when he adds that the agent must also be engaged in the second-order pretense.

But if the reason for engaging in the second-order pretense is to make the first-order pretense more credible — i.e., to conceal it as pretense — how are we to make sense of the act of engaging in that second-order pretense without attributing to the agent the very knowledge that would undermine its aim? It seems that the agent must intend to get himself into the state of engaging in both pretenses for the sake of achieving a psychological end, but he must somehow manage to do this without revealing to himself that this is what he is doing. If the problem with intentionally trying to acquire a belief that one also believes to be false has to do with the fact that (acquiring this kind of) belief is not under control of the will, the problem here is that what one is able to conceal from oneself about what one is doing is not under control of the will either.

One way of bringing out this difficulty is to ask why Darwall thinks second-order deception will be enough to do the trick. If the reason for engaging in the second-order pretense is to make the first-order pretense more credible, then wouldn't we also require a third-order pretense to conceal the nature of the second-order pretense as pretense? And so on, seemingly indefinitely? The self-deception-as-pretense strategy seems just to harbour the very same dynamical problem as the naïve view in a somewhat concealed form. The dynamical problem becomes: How can I act in a motivated and purposive way to get myself into a state of pretense which I do not know is a pretense?[15]

Still, an interesting question lurks here. It seems that Darwall wants to appeal to the idea of an elaborate pretense to substitute in to do the work of belief where he thought a belief didn't belong. But what *is* the difference between belief and pretense? Darwall appeals to an agent's pretending that something is the case where he couldn't make sense of the agent believing that very same thing, so he must be imagining that belief and pretense are at least different in that in some cases where an agent is not in a good position to believe something, the agent might nevertheless be in a good position to pretend that it is so. We do, of course have an ordinary grasp on the difference between belief and pretense: it is that you don't really believe what you merely pretend. But that's clearly not going to do the trick here. And this can seem to be a tricky task: in both belief and pretense one acts in characteristic ways with respect to the proposition in question: one *takes* it to be true in action and in speech. Of course there are other more 'internal' respects in which belief and pretense may be thought to differ. The believer, but not the pretender, has a certain felt conviction in the thing believed; the believer, but not the pretender may use the belief in question as a premise in further reasoning — although the pretender may do this is in a way which is circumscribed by the fantasy. But as the fantasy becomes more elaborate, these distinctions become harder and harder to draw. Someone who is thoroughly taken in by their own fantasy, may indeed come to experience a certain felt conviction concerning the thing in question. Indeed, it seems that the more complete the fantasy is — certainly the more higher-orders of pretense one adds to the fantasy (and it seems Darwall may really need a great many) the more it can seem indistinguishable from belief. Once we have ensured that we are to be barred from thinking of our fantasy as a fantasy by the second-order pretense there seems to be no reason remaining why the deceived person should *not* use as a premise

[15] A version of this objection, it seems to me, works as well on D'Cruz's account, according to which the pretense involved in self-deception is somehow *unwitting*. At first, it might seem that understanding self-deception as unwitting pretense might seem to help with this difficulty a little bit. After all, since the pretense is unwitting, perhaps it comes, as it were, with its aims already concealed, obviating the need to posit an ever-higher-order state of pretense to do the concealing for us. But I worry that understanding unwitting pretense as the kind of metacognitive failure D'Cruz appeals to (failure to keep track of which of one's pretenses are pretenses) might just reinstate the paradox in a different form. It may be true that once one has lost track whether one is pretending the pretense will no longer be transparent in the way that it is if it is witting. But, again, how does one intentionally get into *that* state? Can one, in a motivated and purposive way, make it the case that a particular meta-representational content is only intermittently available? Why is this any less puzzling than trying to act in full knowledge to conceal as pretense a pretense whose aim can only succeed if one is unaware that it is a pretense?

in further reasoning the thing that they are deceived about.

In the literature on psychiatric delusions, there is a parallel debate between 'doxastic' and 'non-doxastic' theorists. Non-doxastic theorists (such as Currie 2000) start by observing that many delusions can fail to exhibit many of the central features of belief: delusions are highly resistant to counter-evidence; they are often highly inferentially circumscribed; they can fail to guide action in the way that belief typically does, and so on. Then they go on to identify some other cognitive state which doesn't typically exhibit these features which belief does, e.g., imagination: it is no strike against my imagining that $p$ how much counter-evidence to $p$ I have encountered; there needn't be any rational integration between my imagining that $p$ and any of other doxastic attitudes that I might have; I don't normally take what I merely imagine to be the case to be a sound basis for action. But, the doxastic theorist then replies, not all delusions exhibit all of these failures at the same time, or even typically. Sometimes patients *do* revise or attempt to explain their delusions when confronted by an interlocutor; sometimes they *do* attempt to integrate the delusion with other things they believe (often with disastrous epistemic consequences); sometimes they *do* allow the delusion to influence their behaviour. And delusions do exhibit other characteristics that beliefs typically have: subjects are highly inclined to assert that $p$ when asked; subjects have a strong felt conviction that $p$. So, the response continues, if what you mean by 'imagining' is something which is only sometimes different from belief in a limited number of ways, but which also significantly overlaps with belief in most central cases, then I am not convinced you have really identified a distinct state.

I am highly sympathetic to the doxastic theorists in this debate. The attempt to try to cast delusions as non-doxastic on the grounds that delusions fail to exhibit some of the 'central' features of beliefs relies on the assumption that there *are* such central features, without which something simply cannot count as a belief. But application of our concept of belief is a highly complex phenomenon: there are many independent phenomenological, agential, and inferential criteria for the application of the belief concept and there appears to be no reason to emphasize any of those criteria at the expense of any others. Certainly when we have examples of phenomena that satisfy different, perhaps non-overlapping, application criteria it would seem at best arbitrary, and at worst dogmatic to insist that there is a difference in kind between them.

The self-deception-as-pretense strategy not only harbours the very same paradoxicality of the naïve view in a somewhat concealed form, insofar as the it requires that there be a difference in kind between belief and complex pretense — to which is bears a great resemblance — we ought to find it unsatisfactory.

## Deflation

Let us next address the family of views about self-deception which try to resolve the surface paradoxicality of the naïve view by denying that that self-deceived agent genuinely intends to being about in herself a belief which she believes to be false. I will work primarily with a single example of such a view, Al Mele's, which has been highly influential.

According to Mele, in cases where I deceive myself, I come to believe something which is false by treating the evidence in a motivationally biased way if this treatment of the evidence is what causes me to form the belief. I never intend to deceive myself, merely to redirect myself away from certain kinds of evidence (Mele 1997). Mele's proposal is, in effect, to deal with the surface level paradoxicality in the naïve description by claiming that at no point does the agent intend to deceive herself, but merely to avoid facing up to certain distressing evidence.

From a certain point of view, intentionally redirecting oneself towards certain evidence and away from other evidence can look as puzzling as intending to deceive myself. Suppose my son is missing and the evidence is amassing that he is lost at sea. I may well have motivationally induced biases that cause me to ignore the evidence pointing to the fact that he is lost at sea, and to focus only on evidence for the proposition that it is not yet certain what happened to him.[16] But in order to be directed away from evidence for the proposition that my son has been lost at sea, I must know what would constitute such evidence, and in order for that, I must know what the thing is that such evidence would be evidence for — I would have to know what would be evidence for my son being lost at sea, and then ignore it precisely because it is such evidence. Then, when asked about what I believe about my son's fate, what I report may be sincere but inaccurate. It seems to me for all the world that the evidence favours the proposition that my son is fine after all, or that it is uncertain

[16]This example is from Williams' 'Deciding to Believe' (1973). I am sympathetic with Williams' position, as espoused there, but I do not follow follow him entirely to the conclusion that believing at will is impossible. Perhaps the real problem with believing at will is that belief in the absence of sufficient evidence is simply not possible. This is perhaps a fairly neutral-ground way of articulating why the dynamical problem of self-deception is a problem. If the evidence available is insufficient to warrant or allow belief, in most cases it will also fail to warrant or allow belief when the agent has *intended* to form the belief. So, it is the fact that belief is constitutively norm-governed that explains the psychological impossibility of believing at will. On this way of going, then, it should be perfectly possible to believe at will *if* so believing were to somehow also constitute evidence for the belief itself. Thus, perhaps I could come to believe that I am able to jump the chasm (Johnston 1988) because I need to jump the chasm *if* having that belief were to increase my chances of jumping the chasm successfully. In that case, the normative requirement that belief be a response to evidence would be satisfied, so the psychological impossibility should be expected to vanish.

We might thus come to think that a very moderate form of voluntarism about belief is compatible with the thought that belief must (in the psychological and the normative sense) be a response to evidence. This moderate form of voluntarism could be, at least in principle, actually confirmed by instances of someone successfully forming such a bootstrapping belief. I don't, myself, know whether it is possible to acquire such a belief about one's own caprioling abilities *in situ*, and I confess that my confidence is not high enough to make me particularly enthusiastic about the prospect of facing an otherwise unjumpable chasm any time in the near future. It is important to note, however, that my prospects for forming such a bootstrapping belief are made *significantly* grimmer if I already believe that I *can't* jump the chasm. If belief is transparent to evidence in the way that we are imagining, this appears to make good sense. If I find myself in a situation where the evidence available to me simply doesn't pronounce one way or another on whether I can jump the chasm, I may be able to tip the scales just enough to make belief possible by believing, viz., by producing some evidence. But, if I also already believe that I *can't* jump the chasm, this is presumably because I take myself to have evidence for *that* proposition, evidence which is unlikely to be outweighed by that which I am able to conjure using my own devices. Insufficient evidence means no belief. Contrary belief, on this view, implies insufficient evidence.

what has happened to him, since my attention has been selectively redirected only to the evidence consistent with the truth of this proposition. But now it seems that we have just returned to the original problem again. I am here, in no uncertain terms, intending to deceive myself about that matter: I am actively eschewing evidence for that proposition *because it is evidence for that proposition.*

This is a very tempting thought — and it would be a serious problem for Mele's view if it entailed that this is how self-deception usually works — but I think it is a little too fast. Call this worry 'the appreciation of the evidence problem' as we shall return to it below. One possible response to this problem is to claim that there is a way of directing oneself away from pertinent evidence which falls short of requiring full purposiveness in doing so. This may be a promising strategy (and indeed it is Mele's), but it is constrained by a consideration pushing in the opposite direction: we do not want it to be that the person's coming to have the false belief is brought about by some sort of a deviant causal chain. If, for example, I was caused to ignore all and only the evidence pointing to the truth of the proposition that my son has been lost at sea by a surgically placed blow to the head, we would have avoided the dynamical problem of self-deception, but only by moving to a case that is clearly not an example of it. Mele thinks he can provide just such a middle way. That is, he claims to be able to provide a way of, say, selectively directing one's attention away from pertinent evidence which is not, in the appropriate sense, carried out by the agent or any sub-system of the agent *with the purpose* of self-deceiving, nor with the purpose of ignoring evidence the force of which one already appreciates. In effect, Mele claims, I can intend to do something which *is* an act of deceiving myself, but I don't intend to do it *as* an act of deceiving myself.

His example is that of Beth, a 12 year-old girl whose father has recently died. She has come to form the belief that her father loved her the most of all his children. She has come to this belief by selectively attending to pleasant memories of her father playing with her alone, and selectively ignoring those memories of her father playing with her brothers. Her evidential selectivity is explained by a motivation — a motivation to attend to pleasant memories over over unpleasant ones — but that motivation is not a motivation to deceive herself.

This case is compelling on its face. Beth's case does appear to be one where she acts intentionally in a readily understandable way that leads her to a self-deceptive belief. The question to ask at this point is whether the general dynamical problem will not arise again *for at least some cases.* Is it plausible that all cases can be assimilated to this model?

The case of Beth has some plausibility because the evidence that she is led by bias to entertain (the comforting memories of her father playing with her) has some psychologically pleasant quality to it which is independent of the conclusion that it warrants; similarly, the evidence that she avoids (the times her father doted favourable attention on her brothers) has an unpleasant quality independent of the conclusion that *it* warrants. That is how we are able to get our 'middle-road' explanation: all we have to assume is the apparently quite innocent thought that Beth is motivated towards entertaining pleasant memories and motivated away from entertaining unpleasant ones. But spelling-out the case reveals that things are not so simple, and that many elements of the case must work together in a very

precise way for Mele to get his result.

Mele's case relies on the thought that the memories that Beth entertains of the father playing with her are intrinsically pleasant to entertain, and that the memories she ignores — of her father playing with her brothers — are intrinsically aversive to entertain. One plausible way to imagine this is to suppose that the memories are memories of experiences which were themselves pleasant or unpleasant. It is a striking and essential feature of memory that it encodes — although undoubtedly quite often in a biased way — the affective contours of what is remembered. Beth's memory of playing with her father is a pleasant memory because it is a memory of an experience which was pleasant. According to this story, the affective quality of the memory traces back to the experience it is a memory of, and is, we might assume, minimally influenced by Beth's present mental configuration. But why is it that her memories of her father playing with her brothers are aversive to her? One quite plausible answer, one which is unavailable to Mele, is to say that she finds entertaining those memories unpleasant *because* they are evidence for the proposition that perhaps her father loved her brothers more than he loved her — or at least, evidence *against* the proposition that her father loved her best. To say this would be for Mele to give up the game, since it would be tantamount to admitting that the appreciation of the evidence problem afflicts the view: Beth intends to divert herself away from the unpleasant entertaining of memories which derive their unpleasant character from the forbidden belief that she has, for which they are evidence.

What we seem to need is for there to be a feature of Beth or her situation as represented in the memory which plausibly collude to make her experience of the thing she now remembers unpleasant. There may be a number of different ways of accomplishing this, but it is perhaps most natural to suppose that Beth, as a younger child, was quite jealous. She never much liked it when her father doted attention on her brothers. Even then she was disposed to feel like her self-esteem, or the quality and permanence of her relationship with her father, was threatened by the existence of appreciably successful competition for his affections. As such, she experienced the observation of her father behaving affectionately with her brothers as unpleasant; the very sight of them together would fill her with dread. If we locate the source of all of this unpleasantness for Beth in the original experience, we may be able to get the result that Mele wants. However, we must not think that Beth finds entertaining just those memories unpleasant because she is jealous *now* — then we come dangerously close to a case where Beth intends to deceive herself. For, to be jealous is to have a complex, highly negatively valenced disposition of evaluation and reaction to evidence for the truth of a *proposition one finds aversive*. But in order to find such evidence aversive, one must know what proposition it would be evidence for. To intentionally avoid such evidence, one would have to intentionally avoid evidence for that proposition because it is evidence for that proposition — this would lead again to the appreciation of the evidence problem.

Let us notice how many things need to align in order for this case to be of the sort that Mele needs it to be. Evidence given through memory is peculiar. If we suppose that it has an aversive character, it can have that character either because it encodes an experience which was itself intrinsically aversive, *or* because, taken as evidence, it is evidence for some further

thing which the subject finds aversive. It is the latter option, I submit, that is more typical of cases where people find evidence aversive. When memory is the mechanism, it seems that a motivationally biased disposition of evidential evaluation could be operating for either of these reasons, but that makes memory, as a source of evidence, rather peculiar.

Memory is one of the only ways that the intrinsic character of an experience can be sustainably represented to an agent over time. Memory, if it is genuine, encodes experiences *as* experiences had by the remembering subject, and to the extent that it has high fidelity, it also encodes to an impressively high degree both cognitive and conative aspects of the experience; it is plausibly the mechanism responsible for connecting time-slices of a person over temporal distance. The degree to which a person is connected with past versions of himself, so to speak, depends on the strength of his memory connections. My sense of myself as a future survivor of the nervous lad stealing his first kiss on the playground that one Autumn afternoon all those years ago depends crucially on the strength of my memory of that experience along both cognitive and conative dimensions. But it is implausible to think that all forms of evidence I might encounter — or for that matter purposively ignore — have this unique feature. And if they do not, then it seems much more plausible to think that, if that evidence is found aversive, then it is only aversive *because* of the aversive character of what it is taken to be evidence for. But in these cases Mele's account seems to face a challenge: If someone only finds evidence (un)pleasant in relation to the (un)pleasantness of what it is evidence for, then her avoidance of the unpleasant bunch is either motivated and purposive or it is not. For it to be a genuine case of self-deception, it seems it must be motivated and purposive. But, then we seem to have a version of the dynamical problem again. For how can she be motivated to avoid evidence if the aversive character of the evidence derives from the aversive character of the thing which it is thought to be evidence for — the very thing which, if she purposively avoids, she is guilty of purposively deceiving herself about?[17]

This problem does not arise with Self-Deception as Omission. According to my account, there is no need for an agent to be *motivated* to avoid particular evidence. The agent may in fact be caused to avoid certain evidence via the operation of a System-1 style affect-regulating mechanism, but that will not necessarily be something that the agent himself can properly be said to do for a reason. Motivation is still playing a role on my account, but it comes in at a later stage, playing a role in maintaining the belief, rather than one in the formation of the belief itself, so there is not need to think that agent has already interpreted the evidence against the belief as threatening.

It is implausible to think that we typically avoid evidence because it is intrinsically

[17]What I take to be a version of this kind of objection to Mele's account has been given by Robert Lockie (2003). However, Lockie takes the objection to be part of an argument for what he calls a 'depth-psychological' account of self-deception, and I take my framing of the objection, and where I intend it to lead us, to be accordingly importantly different. However, I do think Lockie is onto something. I am willing to concede, for example, that we may, because of self-deception and related phenomena, be forced to countenance the existence of a 'dynamic unconscious', but as I argue below, I worry that accounts of self-deception that appeal to such an unconscious get the facts about responsibility wrong.

aversive. Many more typical cases of self-deception do not exhibit the features of Mele's case where many factors must collude for it to show what he takes it to show. Suppose I am in the early stages of a promising and exciting relationship. Now suppose I encounter some evidence: the object of my affection is seen at the ballet with someone else. It would be, I think, at least uncharitable, to impute to me a reaction which finds that minimal description aversive without further specifying that I find it aversive for a rather specific reason: because I take it to be evidence for something which, if true, I would indeed find distressing, namely, that she does not share my hopes of pursuing an exclusive partnership. Why should I find the evidence itself aversive? The person she was with may have been her brother; she may have been invited by an acquaintance who happens to be a local critic and who requires her discerning sense of costume and set design; she may have been mistaken for the Queen of Denmark (on account of her very elegant dress) and brainwashed to attend against her will by fanatical royalists. (I may, of course find this evidence distressing because it is evidence for *other* distressing propositions.) That is not to say that I, or we, do not typically fill in such gaps to facilitate the seamless drawing of the distressing inference; the point is just that in the absence of the inference, there is nothing to find distressing. But in order for the inference to be made, it seems that we must return to the apparently puzzling place where we began.

I therefore do not think Mele's account provides a satisfying general solution to the dynamical problem. At bottom, this is because Mele's view still requires self-deceivers to *act* in a way that is guided by the intrinsically aversive quality of the evidence that, if appreciated, would undermine their self-deception. And I have just questioned whether it is plausible that the evidence should have that aversive character (and thus be available to play that explanatory role) independent of belief about the proposition which the evidence warrants. But this problem is avoided on my view. There is no action the explanation of which requires us to think some evidence has an aversive quality, either intrinsically or derivatively. What has an affective quality (albeit a positive one) is the belief that the agent already finds herself with, and what she does — or rather, omits to do — is explained by that belief and its attendant affect itself. And there is nothing mysterious about why the belief in question has the attendant quality that it does: it is *just more pleasant* to believe that I am not going bald than to believe that I am.

My view does undoubtedly share some features with Mele's view. We both agree that the dynamical problem ought to be taken seriously, and we both make appeal to cold bias in an attempt to do so. But Mele still identifies *an act* of self-deception. If we were so inclined, we could anchor our judgements of responsibility with that act, but it is also that very thing which leads to the appreciation of the evidence problem, which I take to be a version of the dynamical problem. Self-Deception as Omission does not face this problem.

## Fragmentation

One further, and arguably the most popular,[18] way of explaining how someone could so much as manage to deceive himself is to offer up a psychological picture according to which the operations of certain important motivational and belief-forming processes are insulated from each other, and from 'the person as a whole' in a way that would allow the deceived agent to be both the perpetrator and the victim of his deception. One such partitioning strategy which meets both of these demands would be to posit what we might call a *dynamic unconscious*. If we locate both the belief that $p$ *and* the motivation to believe not-$p$ in the dynamic unconscious we seem to be able to resolve both the (putative) static and (more threatening) dynamical problems at a stroke: My belief that not-$p$ is safely compartmentalized away from my belief that $p$, just as my motivation to deceive myself is safely tucked away from my conscious psychic economy. The sense in which the unconscious envisioned by this strategy must be dynamic, and not merely static, is as follows: The unconscious is thought to be one 'part' — perhaps among many — of a person, and must have its own characteristic motivations which are not known to the other parts (or to the 'person as a whole'), and those motivations must be capable of directing unconscious activity in such a way that can conceal things from, or reveal things to, the other parts and can deceive them, or cooperate with them in accordance with its own characteristic aims.

The strategy of postulating fragmentation in psychological subjects as an explanatory hypothesis is an old one. The earliest known instance of the strategy belongs to Plato who, in Book IV of the *Republic* has Socrates argue that the soul has three parts in order to explain the perfectly everyday phenomenon of being attracted to and repulsed by the same object, seemingly at the same time. Plato also uses the picture to explain the possibility of *akrasia*. In the earlier *Protagoras* Socrates argues that *akrasia* is impossible on the grounds that the soul is uniform and pursues only the good. If whatever the soul desires is apparently good, and there is only one desiring component in each soul, it cannot be that anyone ever willingly seeks the bad. The development of Plato's thought away from the view expounded by Socrates in the *Protagoras* led him to reject this line of thought and instead to accept that *akrasia* is possible in some forms. In order to get this, he needed to reject the assumption that each soul contains only one desiring component.

Closer to our own time, and much closer to contemporary thinking about the mind and human behaviour, Freud, of course, was also a proponent of this strategy. Like Plato, he thought taking the psychological subject to have parts, some of which are hidden and whose operation is unknown to the subject, was warranted by the evidence provided by pathological and non-pathological cases alike. In *The Unconscious*, he writes (Freud 1963, 116–117):

> The assumption of the the unconscious is both *necessary* and *legitimate*...because the data of consciousness have a very large number of gaps in them; both in healthy and in sick people psychical acts often occur which can be explained

[18] Views of this kind have been influentially defended by Davidson (2004a) and Pears (1984).

> only by presupposing other acts, of which, nevertheless, consciousness affords no evidence.

Freud, like Plato, thought that a divided subject, containing independent sources of motivation, was a warranted explanatory hypothesis given the facts of ordinary psychological life, and of pathology. Although I do not think that this strategy, in general outline, is in any way implausible, many philosophers have had some misgivings about it. One family of misgivings gets perhaps its most forceful articulation in Jean-Paul Sartre's discussion in 'Being and Nothingness'. A brief discussion of Sartre's criticism of Freud will help us to get clearer about the advantages and disadvantages of a fragmentation approach to self-deception.

**Freud, Sartre, and 'the mind'**

One caveat is appropriate here: Freud, strictly speaking, never discussed anything he called 'self-deception'. However, the clandestine operation of unconscious motivations is obviously a central feature of his thought, and we owe much of our familiarity with such ideas to him. In particular, the tripartite structure of the ego, the id, and the superego — the form of Freudian theory which crossed Sartre, and with which he took issue — seems to readily offer an explanation for how unconscious deception could take place. On a standard reading, the early Freudian posits the id as a reservoir of unconscious instinctual drives, which, so long as they remain there, also remain unsymbolized (we might say unconceptualized), inaccessible to conscious understanding. The way Sartre reads Freud, the ego was to be identified consciousness, and the contents of consciousness are partially regulated by a 'censor', which stood at the boundary between ego and id, determining which elements of primitive psychic instinct are allowed to make it to the level of conscious awareness, and the form (perhaps quite distorted) in which they will be symbolized once they get there. But, Sartre, claims, this leads to a problem (Sartre 1966, 91):

> The censor, in order to carry out its activity with discernment, must know what it is repressing. If we renounce in fact all those metaphors that represent repression as an interaction of blind forces, we are forced to admit that the censor must *choose*, and in order to choose, it must *represent itself*.

The thought seems to be: Sometimes it comes about in me that I have a belief which I take to be true, but which has been caused in my by the discerning operation of the censor. I, as ego, believe to be true what I, as censor, know to be false. But how, Sartre wonders, could this be? Doesn't it seem that in order for the censor to effectively play this role it would have to be in the very same puzzling state that I, unanalyzed into Freudian modules, seemed to have to be in according to the näive view of self-deception? Wouldn't the censor have to know not only the content of what is being allowed into consciousness, and permit it entry based on the projected psychic results of doing so, but also *that it knows the belief that it is bringing about in consciousness to be false*? Sartre, in effect, seems to be claiming

that the dynamical problem would apply no less to the Freudian censor than it would to unanalyzed agents.

On the face of it, this would seem not to be a very plausible reading of Sartre since, even if we grant the surprising assumption that all knowing requires second-order knowing-that-one-knows ('representing' oneself as knowing), this would not yet get us to the dynamical problem. The dynamical problem is supposed to point to a fraught variety of conscious or intentional mental activity: there are constraints on the sorts of things that can be simultaneously brought into a single conscious mind at any given time, and believing something to be false and intending to get yourself on that basis to believe that it is true at the same time seems to plausibly violate those constraints, however we choose to precisely spell them out. But merely having second-order knowledge that it knows, while it may be thought required for the censor to 'apply its activity with discernment', doesn't yet obviously violate the constraint on conscious mental activity that the dynamical problem is based on. After all the censor's first-order knowledge, or second order-knowledge of what it knows, or both, could fail to be conscious.

In order to get to his conclusion that the Freudian has failed to resolve the original puzzles associated with self-deception Sartre seems to have to assume a couple additional things. First, he seems to have to assume that all knowledge (or belief) is conscious. Only if this is the case can we sensibly talk of the violation of constraints on what can happen in a single conscious mind simultaneously. But further, Sartre seems to need to assume that the consciousness of the censor is *my* consciousness. Otherwise, we would end up with a picture where there is a conscious component of my mind which is distinct from me, whose function is to determine via its own conscious mental activity which things are to make it into my, distinct, consciousness, and in what form. Although this view is certainly implausible, as far as constraints on unified conscious mental activity go, it is entirely beside the point, since it involves not one locus of conscious mental activity, but two.

If we grant Sartre these two assumptions, this would get us back, in our terms, to the dynamical problem: the censor would at once have to know the aversive or troubling content repressed in the id, know that it is being allowed into consciousness to be symbolized in a radically different (untrue) form *and* know that it is doing this for the purpose of producing this distorted representation *in a consciousness which is not distinct from its own.*[19]

Of course, if these assumptions were true, I think we could grant Sartre's charge that the Freudian has not solved the puzzles of self-deception. But, neither assumption seems plausible, and the choices that we seem to be forced to if we accept them bear this out. If we accept both, we are led to the dynamical problem. Which assumption is the problem? Although the common consciousness assumption may seem like the more egregious of the two — and it certainly is a serious misreading of Freud — it should actually be seen as

[19]Of course, it won't do, at this point, to attempt to explain how this could be possible by claiming that although the censor has knowledge of what it is doing, this knowledge is unconscious. This is because if the explanation of how repressed knowledge is to remain unconscious requires the operation of a censor, we would seem have to have to postulate a second censor. If the activity of this censor is to remain unconscious, we should have to postulate a third censor, and so on to an infinite vicious regress.

secondary to the much more troubling assumption that all mental activity, certainly all purposive mental activity, must be conscious. Indeed, we were only led to the problem of trying to find *some consciousness* for the censor's knowledge of its own activities to occur in on the assumption that that knowledge had to be conscious in the first place. From the perspective of a Freudian explanation of unconscious deception, the whole point of postulating a dynamic goal-directed mechanism which operates without my awareness for its explanatory value is obviated if we assume that no such thing is possible from the get-go on the basis of the assumption that mentality and consciousness are identical.

From the perspective of modern cognitive science, the view that all mental activity is conscious mental activity seems positively archaic. Sartre's misreading of Freud is relatively well-known, but it has potentially surprising relevance for us. As a response to Freud on the existence and operation of the unconscious Sartre's response is flatly question-begging. And this means that if we are to produce reasons for disfavouring a Freudian-style account, we shall have to look elsewhere.

A Freudian-style 'partitioning' strategy to deal with the surface paradoxes of self-deception takes the apparent paradoxicalilty in the naïve description of the phenomenon to express a kind of constraint on what it is possible to simultaneously consciously have in mind. If we understand the paradoxes in this way the Freudian solution presents itself as a natural, even obvious or flat-footed, response: simply deny that puzzling co-occurring phenomena are both conscious. We needn't adopt any of Freud's substantive views on the contents of the drives, or any of his developmental doctrines, in order to accept what we might call a 'Freudian' solution the paradoxes of self-deception. If we read Sartre's objection to the positing of unconscious processes as a simple refusal to accept that anything genuinely *mental* could fail to be conscious, we should be unanimous in our refusal to follow him there. Indeed, any resistance to an account of self-deception ought not come from skepticism about the existence and operation of unconscious mental processes. In fact, it would surely be a mistake to locate any important current dispute over the validity of the distinction between the elements of the mind which are conscious and those which are unconscious; we know that there are efficacious unconscious mental processes. Perhaps even the vast majority of the mind is like this. Although, of course, the modular mechanisms of cognitive science do have one important major difference with Freudian mechanisms. Freud's mechanisms were, by hypothesis, goal-driven. And while we are certain to be hesitant about saying that there are 'unconscious drives' (cognitive modules) that have the particular goals that Freud characteristically ascribed to some of the contents of the id, we should have no difficulty at all accepting the existence of subpersonal goal-directed cognitive processes as such.

So, the Freudian solution has much to recommend it, and the most obvious and influential objections to it can be seen to be misguided. Why then, do I think it is unsatisfactory? I am quite sympathetic to an account of self-deception that looks like this, and some elements of this view are recognizable in Self-deception as Omission. One of the biggest advantages of the Freudian view as a solution to the surface paradoxes of self-deception is that it leaves intact the ordinary notions of belief and intentional action as they appear in the statement of the naïve view. Or at least it is supposed to. However, even if we grant that it solves, in

a way, the dynamical problem, it does not seem able to make sense of the judgement that self-deceivers are responsible for their self-deception.

The Freudian solution threatens to draw the contours of responsibility in the wrong places. One reason the naïve view might have seemed like an attractive place to start was because it promised to transfer our fairly firm judgements about responsibility from the interpersonal to the intrapersonal domain; since the person who deceives acts intentionally he is responsible for the deception, and since in self-deception the deceiver and the deceived are the same, *that person* is responsible for the self-deception. But on the Freudian view, there is no person that is responsible for the self-deception.[20] The main system, or what we might call the ego, which has the most plausible claim to being responsible for the majority of a person's intentional actions as we normally conceive them (speech, bodily motion etc.) is an innocent victim. The deceiver is merely a type of protective censor. But in what sense can the self-deceived be held responsible for a deception which is perpetrated by an autonomous subsystem of his? It is typical of the censure that we feel towards people who are victims of their own self-deception that we think it involves a kind of flight from anxiety, or lacking the courage required to face the facts of a particular situation. And this is an important facet of our thinking which Self-Deception as Omission is able to capture. But the Freudian can capture neither this, nor the bare judgement that the self-deceiver is responsible. The deceiving censor is merely a liar and is guilty of no such epistemic cowardice. On the other hand, since what is allowed to be symbolized and to enter the ego is regulated by a subsystem entirely independent from it, there is no way the ego could have known better, and it is innocently deceived.

It also important that we ask: What does it mean to say that the censor *intends* — in the very same sense that I can *intend* to go to bed at a decent hour tonight — to deceive the ego? To say that an action was intentional can mean something as robust as the claim that the action is potentially practically-syllogisable and this will necessarily involve imputing both beliefs and desires to whatever it is that we take to have performed the intentional action. It is tempting to claim at this point that beliefs and desires and other propositional attitudes can only properly be ascribed to persons and to claim on those grounds that the censor's actions couldn't possibly be practically-syllogisable and therefore couldn't be intentional. We could, of course, reproduce the complexity required for such attribution in the censor, but this would not only belie its purpose as a mere psychodynamic subsystem, it would introduce at last a truly unacceptable kind of homuncularism. Self-Deception as Omission does not require us to impute all this intentional structure to the subpersonal. Indeed, it is an explicit feature of my view that the mechanisms causally responsible for the production of the self-deceptive belief are sub-intentional.

It is tempting to say, on behalf of the defender of a Freudian-style approach, that the agent is responsible insofar as he acquiesces in the suggestions of the censor. But once we take on board the idea that those suggestions are not to be construed in full-bloodedly

[20] Of course, we could add to such a view the idea that something goes on at the personal level for which the agent is responsible. This is precisely what Self-deception as Omission does.

intentional terms, the daylight between this view and the Self-Deception as Omission starts to disappear. This would not, I think, be a vindication of the Freudian-style view, but rather a repudiation of it. To my knowledge, no one who has defended a strategy of this type has attempted to find a locus of agency in a voluntary omission that occurs after the formation of the self-deceptive belief. And I am inclined to think that this is due precisely to fact that on Freudian-style approaches, the operations of the autonomous subparts of the mind are usually thought of themselves in intentional terms. And insofar as these theorists have been motivated to preserve our responsibility judgements (and it is not clear to what extent this is true), it is possible that they assumed we could do so by appealing to the fully intentional operations of the subsystems. However, I have argued both that we shouldn't construe the systems that way, *and* that it wouldn't, on its own, be enough to recover our responsibility judgements. This should push us towards Self-Deception as Omission.

What the Freudian-style account correctly directs our attention to, and what it shares with my preferred view, is the idea that in order to make sense of how someone can be self-deceived, we need to make reference to mental processes whose operations, as the mechanisms of self-deception, are unknown to them. And, insofar as this was part of its aim, it also rightly attempts to preserve our judgements about responsibility and their connection to intentional action by trying to leave intact the feature of the naïve view according to which the self-deceiver's actions *really* are, in some sense, intentional. But it seems unable to make good on this aim.

## 3.5 Affinities With Fingarette

Before closing this chapter, it is worth noting that my view bears some interesting similarities to a view which does not neatly fit into the threefold classification of views I gave in §3.2. Herbert Fingarette (Fingarette 1969) proposes to understand self-deception in a way which avoids the language of belief and intention altogether. Fingarette's idea is that self-deception is a *failure of engagement*. On his view, we can, whenever we act, choose to integrate that action into our narrative self-understanding, our 'personal identity' as he puts it, *as* someone with a certain character, or moral, religious or cultural commitment or affiliation. But, when I deceive myself, I *refuse to spell-out* the consequences of behaving in a certain way. Someone may drink too much, and problematically, but refuse to engage with an identity as an alcoholic. He disavows this identity to himself by refusing to 'spell-out' to himself how his behaviour *makes* him an alcoholic. Fingarette's key notion here is obviously that of 'spelling-out'. In his usage, it is meant to identify a class of related activities which a subject may engage in to make explicit to himself the way in which his various actions and exercises of skill relate him to the world. The skillful exercise of a capacity need not involve any spelling-out; I can play the violin without articulating to myself (linguistically or otherwise[21]) that

[21]Fingarette explicitly intends the analogy with the properly linguistic act of spelling-out to be suggestive but not perfect since presumably there are non- or proto-linguistic ways of spelling-out to myself what I am doing at the moment.

I am playing the violin, that I am performing any of the actions constitutive of playing the violin, or what this means for me as subject in the world. But I may, *if there is reason to do so*, spell-out any of these things to myself at any point in the course of playing the violin.

So, our alcoholic may drink heavily, and we may think that he is refusing to believe something in the teeth of evidence, and that this requires a motivational explanation. But according to Fingarette, he is simply failing to spell-out to himself the consequences of this mode of engagement with the world for his identity; his behaviour makes him an alcoholic, he just hasn't made this fact plain to himself. On this story, he may believe that he's an alcoholic but there is no need to attempt to explain how he intends to make himself believe something that he thinks is false because his situation is not caused by something he intends to do or actively does at all, but rather by something that he doesn't do, something he *refuses* to do.

Fingarette's account might seem to face a familiar dilemma. A failure to spell-out is either purposive and motivated, or it is not. If it is not, then it does not appear to be case of self-deception at all, but rather a failure of self-knowledge, caused by some standard cognitive limitation, or perhaps by organic malfunction. If I fail to spell something out to myself, even though I have good reason to, I am not guilty of deceiving myself if my failure to do so is the result of being caused to fail to do so by direct neural stimulation.[22] Nor, does it seem, would I be deceiving myself if I simply failed to spell something out to myself because I failed to see that the balance of the reasons that I have actually favour my doing so. I may have reason not to drink the stuff in this glass — perhaps it's gasoline — but I am not guilty of *irrationality* for wanting to drink it if I desire to drink some gin and believe that it is gin on the balance of evidence available to me after a sensibly judicious exercise of my evidence-gathering faculties. Similarly, I do not appear to be guilty of deceiving myself if the balance of reasons favours my spelling something out to myself, but those reasons are not cognitively available to me now.

That Fingarette did not intend to be read in this way is quite plausible, but is complicated somewhat by his deliberate eschewing of what he calls 'cognitive-perceptive' vocabulary (which would include 'belief' and 'intention') to describe self-deception. Nevertheless, I think there is good reason to suppose that Fingarette intended the *refusal* to spell-out to be a motivated purposeful action, and would thus be disinclined to accept the second horn of the dilemma. We would then have a firm case of real self-deception, but it might start to look like we haven't yet explained it — and this is the first horn. What reason would the subject have to refuse to spell-out just this very thing to himself? It is tempting to say that it seems like he would have already had to have spelled-out to himself how spelling-out just this thing to himself is a bad idea; if his refusal is to be for reasons, there has to be a reason to avoid spelling-out, but those reasons are only apparent if one has already done the relevant spelling-out and discovered that it leads down an affectively forbidden path, one which is forbidden because it reveals to the subject a belief which he has a vested interest

[22]Perhaps, if a coherent 'top-down' account could be made out for anosognosia, we would have a real world case of such a thing.

in not facing up to, one that would make trouble for his already existing understanding of himself. Fingarette is absolutely right that seeing how a certain kind of evidence or behaviour threatens one's self-understanding or one's self image is often the basis for self-deception. But if we take seriously the idea that someone's refusal to come to grips with a certain bit of behaviour or evidence is *purposive*, we seem to to arrive at once back at the original paradoxicality: How can I purposefully refuse to spell-out how $\phi$-ing means that not-$p$, unless I already believe that my $\phi$-ing means that not-$p$?

If we are to capture the sense in which self-deception is a kind of flight from distress (very broadly construed) the *reason* for the self-deception must have something to do with the avoidance of that distress. Fingarette insightfully notes that our beliefs about our identities and our relations to the world are among our most guarded, precisely because we have a high degree of motivational investment in their truth; it would be nothing short of a psychological catastrophe for me to find out that my most basic and cherished beliefs of this kind are false. How, on Fingarette's view, do we come to this self-knowledge? Is it *only* through spelling-out? If so, there is room to say that the person who refuses to spell-out needn't have already done any of the relevant spelling-out, but this would open up the possibility that the kind of self-knowledge in question could be achieved — and threatened — in ways other than by spelling-out. But that is just to say that deceiving oneself and refusing to spell-out something which threatens one's self-understanding are not the same thing: if my self-understanding can be threatened by things other than spelling-out, I can self-deceptively avoid that threat by refusing to do whatever that other thing is. But, if the kind of self-knowledge in question can only be achieved or threatened by way of spelling-out, it seems that in the case we are imagining the purposive self-deceiver must have already done some of the relevant spelling-out to have so much as an inkling that he should spell-out no further.

Two responses to this come to mind: The first is to question whether refusing to spell something out for a reason really requires that the subject have already done some of the relevant spelling out (or indeed to have any particularly hard-won self-knowledge); and the second is to question whether having done this spelling-out means that the subject has already acquired the belief that he was trying to avoid having by refusing to spell-out. What is crucial, however, is that the subject need not have any affectively forbidden belief in order for his refusal to be motivated, and this pushes us towards Self-Deception as Omission.

The objection thus far envisioned presupposes that there is some knowledge or belief that the subject must have if we are to make sense of his refusal to spell-out as purposive in the right way to call it a kind of self-deception. In particular, the objection presupposes that the subject must believe precisely that which spelling-out that which he refuses to spell-out would reveal to him. But this isn't true. Self-deception must have a purpose and the self-deceptive process must therefore not be completely brute or blind, but it does not follow from this that it must be fully luminous to the subject. Our alcoholic may refuse to spell-out to himself what his behaviour means for his identity, but this needn't be motivated by a belief about what it would mean for his identity if he were to proceed. The belief that

he is *not* an alcoholic could have come about as the result of the operation of a DARM.[23] All we need to assume to motivate his refusal to spell-out the consequences of his behaviour is that the positive affective consequences of believing as he already does motivate him to forswear further investigation (spelling-out, after all, is a form of investigation) and to acquiesce. (Whether we should say therefore that he hasn't spelled-out the consequences of his behaviour — because so doing would lead him to a belief that he need not have — or that he has spelled-out just enough to learn that he is happy where he is, epistemically speaking, seems to be a matter of free choice). Refusing to spell-out may be nothing more than the acquiescence which has been central to my account all along. The kind of failure of engagement Fingarette has in mind is very similar to the culpable failure to do what is necessary to ensure that one is appropriately responsive to certain reasons for belief that is central to Self-deception as Omission.

[23]Once again, in the following chapter I will consider a logically weaker version of Self-Deception as Omission where the requirement that the belief come about as the result of the operation of a DARM is dropped.

# Chapter 4

# Self-Deception and Delusion

## 4.1 Introduction

In this chapter I raise a somewhat uncomfortable question: Are at least some delusional subjects *responsible* for their delusions? The question strikes us as uncomfortable at least in part because we think the answer is just pretty clearly 'no'. Nevertheless, I will argue that at least some delusional subjects are responsible for their delusions. My argument will be as follows: When we consider the dynamics of self-deception — as I have argued we should understand it in the previous chapter — we will see that there is enough overlap between them to ground the judgement that self-deception is implicated in the formation and maintenance of at least some delusions. Since, as I have said, we typically think self-deceivers are responsible, and my account captures the sense in which this is correct. I then argue that, according to my account, at least some delusional subjects are self-deceived. Importantly, I believe that this can be shown to be the case without leading us to the judgement that delusional subjects are blameworthy for their delusions. In order to thread this line, I will appeal to the distinction between what I will call 'attributability' (roughly[1] following Shoemaker (2011) and Watson (2004)) and blameworthiness. I will argue that while self-deceivers are typically responsible both in the sense that their self-deception is attributable to them and in the sense that they are blameworthy, delusional subjects, even when they are self-deceived, are typically only responsible in the sense that their delusions are attributable to them. Why this should be so will be made clear by consideration of the details of my own view of self-deception, as well as the details of the delusions which I consider. In the process I will elaborate some of the features of Self-Deception as Omission, and will give a revised (indeed, logically weaker) version of the view which highlights the dynamics I wish to focus on.

A little bit more about the significance of our question: lying behind the seemingly ordinary idea that delusional subjects are not responsible is the idea that delusions are somehow

[1]But only roughly. I explain how my use of this term — and the distinction I use it to mark — differs from Shoemaker's below.

beyond the scope of ordinary interpersonal understanding. Karl Jaspers (Jaspers 2007) famously distinguished this kind of understanding from what he called 'explanation'. Jaspers was aware that psychiatry was partly a natural science, and partly a human science. Explanation is what natural science does: it uses objective empirical methods to elucidate causal structures. Understanding, on the other hand, is unique to human science, and uses ordinary interpersonal imagination and other 'subjective' methods to appreciate the experiences of subjects 'from the inside' (Kendler and Campbell 2014). Jaspers nevertheless found that understanding could sometimes break down in the face of more extreme symptoms. The prevailing ethic in contemporary medical psychology seems to agree with him. The idea is to regard patients suffering extreme symptoms as deserving of compassionate treatment, but also as nevertheless, at some ultimate level, perhaps *beyond understanding* — or as Jaspers himself put it, 'un-understandable'. I will not (and cannot) argue that understanding does not break down in the face of some extreme conditions, but it is my view that we should push the boundaries of such understanding as far as they can go in the hopes of coming to grips with how best to understand, in ordinary humanistic terms, what is going on in certain forms of mental illness. I will return to some of the consequences of my argument for the understandability of delusions in §4.5.

So, the uncomfortable nature of our question belies a commitment to extending ordinary human understanding — and indeed, the boundaries of the moral community — as inclusively as we can. For non-experts, our understanding of delusions depends on a highly elaborated medical practice to which we are largely outsiders. And I wish to take seriously the critical idea that practices such as institutionalized medicine and the knowledge which they enable often conceal dynamics of unequal power (Foucault 1969).[2] This behooves us to be sensitive to the tacitly normative aspects of the explanatory categories appealed to by such practices (categories such as *delusional*), and the effects that such categorization may have on those who are subject to it. Whether someone who is suffering from delusions is — and whether it is appropriate *eo ipso* that they should be made to *feel like* — a non-agent, a passive sufferer, or someone who is generally non-responsible, are philosophical questions, and answers to them should not be implicitly imported along with the very idea of a delusion. This discussion is an attempt to provide a philosophically sound way of broaching these questions, and to temper the temptation to give too-easy answers to them.

So, I think our question is important. As I said, to go about answering it I will argue that there is overlap between delusion and self-deception. More precisely, I will argue self-deception can play a role in the formation and maintenance of delusions. But as I said (and I hope to be able to illuminate why this ought to be so) we typically judge self-deceivers responsible for their self-deception. We are also, as I said, pulled towards the claim that delusional subjects are not responsible for their delusions. Taken together with the thesis

[2]Foucault located such dynamics within what he called 'discourses', which are 'ways of constituting knowledge, together with the social practices, forms of subjectivity and power relations which inhere in such knowledges and relations between them' (Weedon 1987, 108). I do not wish to problematize the knowledge that discourses enable (as Foucault did) but merely to draw attention to the tacitly normative aspects of certain practices of categorization.

that I want to argue for, this suggests the following triad:

1. Delusional subjects are *not* responsible for their delusions
2. Self-deceived subjects *are* responsible for being self-deceived
3. There is overlap between self-deception and delusion

If (1) and (2) are read as generics (and they certainly should not be read as universal generalizations), then the triad is not, strictly-speaking, inconsistent. But it points to the need to say something about how we should think of the identified cases of overlap with respect to responsibility. Are they, in this respect, more self-deception-like or more like typical delusions? I hope to be able to make clear, by way of appeal to my account of self-deception, why we should go for the former and not the latter. With that in mind, let us turn to my account of self-deception.

## 4.2 Self-Deception as Omission (Again)

As I have just argued in the previous chapter, Self-Deception as Omission is preferable to many of the other views on offer in the literature on self-deception. I want now to consider a modification of the views discussed in the previous chapter, one which can be motivated in much the same way. To begin, let's consider an example: [3]

> **A:** A is an academic who is self-deceived about the quality of his own work. A is unhesitant about advertising what he takes to be his own brilliance to others, but it is clear to his colleagues and everyone familiar with his work that the work is flimsy. Nonetheless, A badgers OUP to put out a volume of his collected papers. He avoids situations where he might have to confront his work's obvious shortcomings, and when he does encounter criticism he dismisses it as jealous or vindictive. It is clear that A longs deeply for the respect and admiration of his colleagues, but it is equally clear that his pursuit of it is self-undermining.

What is the right way to describe what is going on with A? Recall that a natural place to begin, and the view which most philosophical debate about self-deception takes as a starting point, is to think of self-deception as the intrapersonal analogue of ordinary other-deception. I've called this view 'the naïve view':

**The naïve view of self-deception:** A is self-deceived that $p$ just in case A believes that $p$ and A has acted intentionally so as to cause A to believe not-$p$

[3] This example is adapted from Doggett (2012).

According to the naïve view, A believes that his work is flimsy, and he has somehow managed to act so as to cause himself, on that basis, to come to believe that his work is not flimsy. There is a way in which this captures the phenomenon. A seems to believe both things, and the more comforting belief seems to be a defensive response to the more sobering one.

According to my own view, there is no single act which is *the act of self-deception*. This is much is familiar from the previous chapter. However, in the previous chapter I was keen to give a plausible story concerning how the self-deceptive belief might come about in the first place. (It is here that I appealed to dual-process theory and to what I called doxastic affect-regulating mechanisms.) However, I now want to abstract away from the process of belief formation altogether. It will be recalled that according to my view, the node of intentional agency for which the self-deceiver is responsible comes in long after the belief is formed anyway. It is here that we find the operation of motivational factors that interfere with the subject's ability to bring her beliefs in line with the available evidence. For this reason, it is not always crucial to self-deception how the self-deceptive belief comes about. Often what is crucial is how that belief is maintained. The view I wish consider now says:

**Self-deception as Omission*** : An agent is self-deceived that $p$ if she believes $p$ and intentionally omits to seek, recognize, or appreciate externally available evidence for not-$p$, for reasons which ultimately relate to her desire that $p$ be true, in a way which enables the maintenance of the belief that $p$.[4]

Severing the connection between some distinctively self-deceptive process of belief formation, and the resulting self-deceptive state provides us with a shortcut for avoiding the dynamical problem. The view also continues to capture — and I will elaborate more on how this is so — the sense in which self-deceivers are responsible: it describes a distinctive kind of motivated epistemic failure which (as we will see shortly) can be grounds for at least a couple of varieties of responsibility judgements.

The revised view also captures what is going on with A. From the vignette we have seen, we don't know how A came to his belief about the quality of his work, but on my view that doesn't much matter. A's belief in the quality of his work may have been well-founded at one point. Perhaps he used to be a big fish in a small pond, outperforming other undergraduates at the small state school where he studied, but as he advanced through his career the abilities of those he was surrounded by rose consistently, while his remained stagnant. Or maybe he was never really cut out for academic work and his belief was formed directly and unconsciously as a way of dealing with the stresses of academic life.[5] According

[4]Note that the view is stated as a sufficient condition. Because the satisfaction of this condition can be sufficient for something to count as self-deception, the constitutive connection thesis (see below) must be false. But that is not to say that there *couldn't* in principle be cases of self-deception which do not satisfy the condition given here. (It would be a separate question, however, whether such cases and the view meant to capture them would run afoul of the dynamical problem.)

[5]There is nothing about the revised view which *rules out* the self-deceptive belief having been formed as the result of the operation of a DARM.

to Self-deception as Omission*, it doesn't much matter. What matters is that *now* the belief is manifestly defeated by evidence which is readily available, but A is impervious to that evidence because he *prefers* to continue to believe as he does — that is, he has a desire that $p$ be true — and this motivates him to forswear looking into the matter any further. This is why he avoids confrontations with others in his field. He omits to do whatever it is precisely that the epistemic norms say that he ought to do in order to bring his belief into proper conformity with the evidence.

What does it mean for the agent's reasons for omitting to seek, recognize, or appreciate evidence to *relate* to her *desire* that $p$ be true? I will say that a subject is *emotionally entangled* with a proposition $p$ if she is liable to satisfaction or dissatisfaction when $p$ is believed to be true or false. To *desire* that $p$ be true is for satisfaction to accompany the belief in $p$ and dissatisfaction the belief in not-$p$ (as opposed to the other way around). Of course, the belief that $p$ does not formally satisfy the desire that $p$ be true. Rather, the formal object of that desire is the truth of $p$ itself. So the satisfaction and dissatisfaction in question for emotional entanglement is not formal, nor is it experiential.[6] It is, we could say, *representational*. It is the kind of satisfaction or dissatisfaction that obtains when the subject's take on the way the world is more or less closely approximates the way the subject thinks the world *ought* to be. This is obviously a matter with motivational efficacy, but it can be so without having a readily identifiable experiential component.

There is admittedly something metaphorical in talk of entanglement, but this expression captures something about the way in which the interaction between, and layering of, desires can produce a complicated web-like structure. There are many ways in which I may desire the truth of some proposition. I might desire it to be true for its own sake (such as I might desire to have a good relationship with someone); I might desire it to be true for the sake of something else (such as I might desire my car to function well); I might desire that something be the case *rather than something else* (such as I might desire my partner's infidelity as an explanation for her growing distance over my own emotional unavailability). These are all things I can be self-deceived about because they are things in the truth of which I can

[6] I use the terms 'satisfaction' and 'dissatisfaction' deliberately to avoid commitment to the idea that there must be positively or negatively valenced *experiences* accompanying the subject's belief. However, there no doubt will be cases where the subject will *experience* satisfaction or believing $p$ or will experience something like *distress*.

manifest an emotional interest.[7] [8]

The revised view departs from the naïve view more explicitly in one very crucial way by finding what is distinctive about self-deception, not in the process of belief formation, but in the dynamics of belief maintenance. It is usually taken for granted in the self-deception literature that nothing should count as a self-deceptive state that doesn't arise from some distinctively self-deceptive process, that there is a constitutive connection between self-deception as a process, and self-deception as the product of that process. Call this thesis the *constitutive connection thesis*. Now, if the constitutive connection thesis is true, it is clear what the object of philosophical analysis ought to be: If what *makes* something self-deception is how it comes about, we had better figure out how it comes about. Of course, there is a trivial reading on which the constitutive connection thesis is true: every particular that comes about as the result of a process is constitutively connected with that process. You don't get goulash by baking pie. However, with self-deception, the connection between process and product is thought to be more intimate than that. The fact that some process was a process of deceiving oneself confers on the resulting state the status of self-deception. Suppose that state is a belief. There are lots of ways to get a belief, many of which are epistemically respectable. Only one way (or some privileged set of ways) of getting that belief is a self-deceptive way. Even though whatever way the goulash comes about will trivially be a goulash-making process, it doesn't much seem to matter (purist

[7]Theorists of self-deception disagree about what form the subject's emotional interest must take. For Mele (2006, 1997), for example, what it is to have an emotional interest in $p$'s being true is to take the error of mistakenly believing not-$p$ when $p$ is in fact the case to be more costly than the error of taking $p$ to be the case when in fact not-$p$ is the case, where that preference is itself to be understood in terms of a motivational bias. So, on this view, I might prefer to believe that I am not going bald to believing that I am going bald because believing that I am going bald if I am not would cause me great distress, much more than the distress that I would feel if I mistakenly believed that I wasn't going bald when I was. Barnes is more explicit about self-deception's anxiety-reduction function. She says 'When a person is anxious that not-$q$, the person (1) is uncertain whether $q$ or not-$q$ and (2) desires that $q$' (1997, 39). Self-deception reduces the person's anxiety by resolving the question of whether $q$ in the appropriate direction. Although my formulation is much closer to Barnes, as far as I can tell, my use of the idea of 'desiring that $p$ be true' is consistent with both of these ways of thinking about self-deceptive motivation. That is, having the motivated error-preferences that Mele is pointing to is as much a matter of being emotionally entangled (in my sense) as anxiously desiring that $p$ be true.

[8]It is also worth noting that both Barnes and Mele are keen to be able to handle cases of 'twisted' self-deception, i.e., cases where the subject self-deceptively believes something he wants *not* to be true. My way of putting things can also handle these cases once we distinguish between wanting something to be true in the ordinary sense and desiring it to be true in the sense that I mean it here. A subject may desire (in my sense) for something to be true (we may say) *masochistically*. That is, he may be liable to satisfaction at believing that $p$ where $p$ is something that is bad for him, something he, in the ordinary sense, doesn't want (or wants the opposite of). Believing that $p$ closes the gap between the way he takes the world to be and the way he thinks the world *ought to be*, but the way that the world ought to be from his point of view is bad, (say, hedonically bad) for him. There are also cases (like those that Barnes considers in detail) where explanation selection is contrastive. I may self-deceptively choose an explanation that I want not to be true (in the ordinary sense) when the only other alternative is something I disprefer. This would thus be a case of desiring the first explanation to be true because if I believe it the world as I take it to be will more closely approximate the way I think it ought to be *relative to the only available alternative.*

intuitions about goulash with science-fiction pedigree notwithstanding) from the standpoint of assessing whether something *is* goulash which *particular* process resulted in it. This doesn't seem to be true for self-deception. Or so the thought goes.

My view involves the denial of the constitutive connection thesis. The way I wish to cash this out it is to appeal somewhat more perspicuously to the distinction between belief formation and belief maintenance. We ought not to think that a perfectly clear *temporal* line can be drawn to distinguish between processes of belief formation and processes of belief maintenance: When is the belief formation process over, and when does the process of belief maintenance begin? I will say that a process (or part of a process)[9] is one of belief formation if the belief counterfactually depends on it for the agent's credence in it to increase; and that a process is one of belief maintenance if the belief counterfactually depends on it for the agent's credence in it not to decrease[10]. On this way of thinking about it, some processes will (relative to a given belief) clearly be processes of belief formation (such as, with respect to the belief that it is raining, the perceptual experience of seeing the rain outside my window); others will clearly be processes of belief maintenance (such as, with respect to the belief that it is raining, not encountering any evidence to the contrary in the meantime); and a great deal will be both (such as, looking again and seeing that it is still raining, as opposed to seeing that it is not).

As I suggested could be the case with A, I think there are cases of self-deception where one initially had good evidence for what one believes, but where the evidential situation has since changed, and this change has gone unnoticed for some motivationally biased reasons. To give another example, suppose I believe that I am popular with the kids at school. Maybe I *was* popular with the kids, but kids are fickle, and they have since turned on me. It seems to me that I might be self-deceived if the reason that I continue to believe as I do is because I am impervious to the manifestly available evidence on account of my preference for continuing to believe as I do. We can suppose that once I have reached the point where the kids have turned on me, my credence in the belief that I am popular is not increasing and *eo ipso* it doesn't counterfactually depend on me doing or not doing anything in order to increase. But my belief does counterfactually depend on my doing something — or more precisely, not doing something — in order for my credence not to decrease. The evidence is manifestly there, and confronted with such evidence a rational agent would revise her beliefs. What's going on with me? I'm self-deceived! I am intentionally omitting to do what is necessary to bring my beliefs in line with the evidence.

What is the nature of this intentional omission? One might wonder the following: If all the agent is doing is maintaining her belief — especially if it is done via omission — in what way is self-deception an intentional phenomenon? For it to be intentional, wouldn't the agent have to be *knowingly* maintaining her belief against available evidence? Doesn't Self-deception as Omission* face a revised version of the dynamical problem? This worry is actually two distinct worries, and I take them both in turn.[11]

[9] I add this qualification to avoid having to individuate processes, and will omit it from here on.

[10] Modulo, should there be such a thing, natural credence extinction.

[11] The worry is obviously related to the worry from §2.3, above, but the connection here with the dynamical

The first worry is that my view won't be able to capture what is intentional about self-deception without facing a version of the dynamical problem. The problem is thought to arise because for something to be self-deception, not only does *something* about it have to be intentional, but it seems that the violation of epistemic norms — the irrationality itself — has to be somehow intentional. This, I take it, is what makes this worry seem like a version of the dynamical problem.

Now, of course, belief maintenance that flies in the face of manifest evidence to the contrary cannot be fully knowing. But it need not be in order to be fully intentional. Here I wish to mimic a move which Al Mele makes in giving his account (1997) of self-deception, and which should be familiar from above. Mele distinguishes, in effect, between intending something *de re* and intending it *de dicto*. According to Mele's view the self-deceiver intends to do something which is an act of deceiving herself without intending to deceive herself as such.[12] This, Mele thinks, recovers what is intentional about self-deception, all the while deflating it to avoid the dynamical problem.[13] I think this is precisely the right move to make *after* we have given up on the constitutive connection thesis. Once we are talking merely about belief maintenance, and not belief formation, the imagined challenge for my view is not to give a psychologically coherent account of some process, but rather to recover a node of intentional agency which is also recognizably an epistemic failure. The answer to this challenge is pretty straightforward on my view: the agent intentionally (*de re*) omits to seek, recognize, or appreciate externally available evidence for reasons that are motivationally biased, even if she does not intend to do these things as such. So, after it has become clear to everyone else that I am no longer popular, I may simply *do nothing* by way of further investigation into the matter and thus continue to believe as I do. The omission may, of course, not be total. It could be that on occasion I do *encounter* evidence but I omit to *put it together* or to *engage with it as evidence*.[14] So long as my lack of further effective epistemic engagement is motivated by a desire that it be true that I am popular, then I will be guilty of self-deception according to my view. And of course this needn't be a one-off affair. My desire that a certain proposition be true may cause me to forgo epistemic engagement on many separate occasions.

The second worry here has to do more directly with my appeal to omissions. It's not true in general that every time I omit to do something, I do so intentionally. If someone in the next room requires aid, but I don't know it, then it seems I don't intentionally omit to aid them if I go on reading my book. But this is precisely the sort of knowledge which is denied to self-deceivers on pain of falling into the dynamical problem. My failure to seek, or my failure to engage with, evidence against what I believe can't be motivated by *knowledge* that it is evidence against what I believe. But the norms that the self-deceiver violates are, in the first instance, epistemic norms, whereas the norm that requires me to render

problem is more explicit and so it deserves a separate treatment.

[12]If Mele would be willing to understand his view as including cases of culpable belief maintenance, our views would be very closely related.

[13]I am not sure that Mele succeeds in avoiding the dynamical problem, for reasons mentioned previously.

[14]Here again we see the affinity with Fingarette's (1969) view.

aid is a moral norm. Moral norms may fail to apply to agents who don't have the right knowledge, assuming the agent is not culpable for her ignorance itself — this seems to be a version of ought-implies-can. But epistemic norms can't be wriggled out of in the same way, especially if they are norms that require an agent to form a particular belief (against a background of evidence and other beliefs). Epistemic norms say how one ought to conform one's beliefs to evidence, or to one's other beliefs, and it is no violation of ought-implies-can that the agent not already have the target belief. If it were, then it would never be epistemically required that anyone form any belief that they do not already hold, no matter how strong the evidence. Ignorance itself can't be – at least not in the same straightforward way – grounds for claiming that an epistemic norm does not apply as it can be with moral norms. So, while it is plausible that some knowledge is required for an omission to count as a violation of a moral norm, it is not plausible in the same way for omissions which are violations of epistemic norms.

What I want to claim next is that the self-deceiver's motivated epistemic failure is a form of mental agency. We interpret him as having some motivations — wanting to believe well of himself, wanting acclaim in the profession, and so on — and those motivations underlie his failure to bring his belief into conformity with the evidence that is available to him. The state that he thus ends up with is, I will say, *a manifestation of his will*. Allow me to elaborate this by distinguishing two different kinds of responsibility.

## What Kind of Responsibility?

To make the distinction I want to make, let's consider a very simple example. Suppose you step on my foot. Naturally, perhaps, I may want to blame you. But first things first. First: are you a *candidate* for moral assessment? Are you a member of the moral community towards whom attitudes like praise and blame are ever appropriately directed? One way to get at this is to ask: Are you a normal adult human being who can recognize and respond to reasons for action? If you're a child, or a paramecium, or — as we too often say — if you're *insane*, you don't have that capacity and we say you are *exempt* from assessment altogether. (Obviously there are some forms of insanity which ground exemptions of this kind. I don't think having delusions, on its own, is one such form. What it means to be 'insane' in this sense is, in part, the topic of current discussion.)

Suppose you're the right kind of creature with the right kinds of capacities to be a candidate for moral assessment. Still, that doesn't settle the question of whether you are blameworthy. We must now ask whether you are *excused*. There are at least two varieties of excuses:

1. *Strong* excuses work by undermining the agent's ownership of the state or action itself. If the action isn't *yours* in the right way you are not blameworthy for it. So, if you stepped on my foot because you were shoved by a passerby, you're not blameworthy because, strictly speaking, stepping on my foot was not something that you yourself *did*.

2. *Weak* excuses block the step from an agent's ownership of the state or action to blameworthiness. If you have a blind spot, and my foot happened to be in it while you were trotting on your merry way, you are not blameworthy. But it's not because you are exempt, nor is it because you failed to act. You act intentionally in stepping, and are responding to reasons (we may suppose), and the action is *yours*. But, you are not blameworthy because of your ignorance. Note that the typical way in which this works is by demonstrating that you did not display a malicious (or, say, negligent) quality of will.

The distinction between exemption and the two kinds of excuses ought to be fairly familiar from ordinary legal reasoning. I must be indicted before charges can be brought against me in a court (this analogous to finding that I am not exempt), and once I am there I can plead not guilty either on account of having not in fact done the thing in question (strong excuse), or on account of having done it in a non-culpable way[15] (weak excuse). Relevant to this, of course, is indeed the quality of my will. Whether I am guilty of malevolence, negligence, or excused altogether will depend on what I believed and what I desired at the time of my action.

We can now define two different kinds of responsibility. The first is:

**Attributability:** An action or state is attributable to an agent iff that agent is neither exempt from the sort of assessment appropriate for that action or state nor strongly excused from such assessment.

Attributability is a way of marking that at least two hurdles have been cleared: you're not exempt, and you're not strongly excused. If you step on my foot because of your blind spot, we can get at least this far. Blameworthiness goes further.

**Blameworthiness:** An agent is blameworthy for an action or state only if that state or action is attributable to her (she is not exempt from assessment and is not strongly excused) and is not weakly excused.

So, only if your action is attributable to you, and you have no excuses that justify the performance of it, is it appropriate for me to blame you. The kind of blameworthiness that I have in mind here is perhaps best thought of as a kind of liability. To be blameworthy in this sense is for a range of reactions of what we might call 'holding to account' to be

[15]The analogy is limited in the following way: There is a difference between what is standardly called 'excuse' and justification. If an agent can produce either, she can be shown to have avoided culpable wrongdoing. An action is justified if, all things considered, it was not wrong (self-defense); an agent is excused if something undermines her responsibility (acting under hypnosis or duress). Either could be presented as a defense against criminal charges, but weak excuses, as I intend them to be, only comprise the latter.

appropriate.[16] These reactions include the Strawsonian reactive attitudes, such as resentment and the withholding of good will, but also things such as demands for compensation, material or otherwise. What unites these reactions of holding to account is that they demand of the offending party a response, the appropriate response to which in turn is forgiveness. The simplest case of blaming someone in this sense is perhaps finding them blameworthy, demanding an apology and withholding good will until it is given. The appropriate response to a sincere apology is forgiveness and a repair of relations. The *compensatory* nature of the demands of accountability which are characteristic of this kind of blame thus distinguish it from punishment, which is retributive.[17]

My use of the term 'attributability' is closely related to that of David Shoemaker (2011) and Gary Watson (2004). For Shoemaker and Watson, judgements of attributability are also grounds for aretaic, or characterological, assessments of agents. My use of the term is in accordance with their use in this respect. So, not only is attributability a logically necessary condition on blameworthiness, it is also in its own right typically grounds for a distinctive kind of assessment. When an agent 'owns' a state or action in the right way, it is expressive of her will in the sense that she thereby reveals to us something of her deep self: perhaps something about her desires and motivations; her perspective on life and on herself; or her characteristic patterns of thought, action, and evaluation.

This can be brought out in connection with the two different kinds of excuses. Since strong excuses work by undermining attributability itself, we should expect that when someone is strongly excused we find that there are no grounds for assessing him aretaically. And this is what we find. If you step on my foot because you were pushed, you don't thereby disclose yourself to me. On the other hand, if you are merely weakly excused you might not be an appropriate target for blame, but I may nevertheless learn that you are clumsy.[18] Sometimes, in addition to blocking the step from attributability to blameworthiness, a weak excuse will

[16]It is worth noting, however, that whereas Shoemaker contrasts attributability with what he calls 'accountability' (as well as a third notion, 'answerability'), my understanding of blameworthiness should not be identified with Shoemakerian accountability. For Shoemaker, accountability has specifically to do with violating relationship-defining demands (which play no role in my discussion).

[17]So there ought still to be a considerable gap between someone's being blameworthy, and the truth of the claim that we should actually *punish* them. Even if someone is found fully blameworthy for an action attributable to them it may still be a wide open question what the right kind of response to their wrongdoing is. Indeed, it is compatible with my way of thinking about moral responsibility that punishment is seldom, if ever, justified or appropriate. We can see this if we imagine adopting a flat-footed consequentialist justification for punishment. What if it turned out that punishment was an ineffective deterrent and a poor means of personal rehabilitation? It wouldn't follow from this that no one's actions were ever properly attributable to them, nor that they weren't blameworthy for those which were bad. It would simply mean that punishment would not be the response justified by those facts. And it seems immensely plausible to me that if anything would make punishment inappropriate, it would be the fact that the wrongdoer already suffers great enough misfortune that further punishment would take on a perverse character. This is just to say that the question guiding this chapter is certainly not the question of whether some delusional subjects should be *punished* for their delusions.

[18]Whether you are excused for being *that way* (or whether it is countervailed by another aretaic assessment) is yet another question which will arise again in connection with delusional subjects below.

also provide grounds for a countervailing aretaic assessment. If I learn that you pushed me to save me from being hit by oncoming traffic, not only are you excused by demonstrating that the quality of your will was not malicious, you show yourself to be acting virtuously, in a way that merits praise. I will return to this function of weak excuses below.

I hope that it is reasonably clear how, according to my view of self-deception, self-deceivers are attributability-responsible for their self-deception, and that they are (typically at least) also blameworthy. The self-deceiver seems to violate an epistemic norm, and so we can begin anew an inquiry parallel to the questions asked when we inquired about whether you were blameworthy for stepping on my foot. Let us consider A. A has somehow come to the belief that his work is not flimsy. But it is manifest that this is not the case. He persists in his fantasy nevertheless. According to my view, this is because he omits to do what is necessary to bring his belief in line with the available evidence because he has a desire that his work not be flimsy. A is not exempt. (In general, it seems self-deceivers are not exempt; no creature without the capacities to be a candidate for moral assessment generally could be the subject of self-deception.) Is A strongly excused? Strong excuses work by showing that the action or state did not 'belong' to the agent in the right way, that it was not an expression of his will. Of course, it is possible for someone very much like A to act, and think, and speak like A and yet to be strongly excused. If A were being controlled remotely via a chip implanted in his brain perhaps he would be strongly excused. But as we are imagining him, A is engaged in a kind of fantasy which serves an important psychological function for him (though he is almost certainly unaware of it), and it reflects, on account of the motivation which my account attributes to him, his desire that things *be a certain way,* a way which they manifestly are not, and from which he has insulated himself. He thus is the owner of his self-deceptive omission(s), he does manifest his will in the process, and, importantly, he discloses himself and is an appropriate target for aretaic assessment.

Self-deceivers often elicit judgements of frustration, pity, and even contempt. These judgements first get a foothold at the level of attributability because they are appropriate responses to someone who has displayed the qualities of character that A has, viz., injudiciousness, vanity, and even cowardice. The only thing which remains to be determined is whether A might have a weak excuse that could insulate him from blame or potentially provide grounds for a countervailing aretaic assessment. But as far as we can tell – and as seems to be the case for self-deceivers quite generally – there is no excuse that A can appeal to. Weak excuses work by showing that the agent did not manifest a malicious or negligent quality of will, but A *does* manifest (at least) a negligent quality of will. Indeed, in self-deception we see the marriage of both epistemic and volitional defects combining to make for this negligence.[19] In willing something to be the case which is manifestly false, A

[19]There are some habits of mind which are epistemic vices only because they are accompanied by indolence. For example, perhaps all of us are subject to the availability heuristic, or liable to commit the base rate fallacy, and various other System 1 cold biases. What separates those of us who allow the errors characteristic of those biases to persistently take hold and those who do not is some degree of epistemic vigilance, which is effortful. For example, consider once again Kahneman's example from above (Kahneman 2011, 44):

both shows the epistemic vice of injudiciousness and is engaged in a flight from anxiety. This combination of epistemic and volitional failures strikes me as distinctive of motivated irrationality. Doing one's epistemic duty often requires a steadier will than the agent possesses, and this failure can manifest itself, on my view, as a motivated failure to seek, recognize, or appreciate evidence. Below I will discuss a case where a manifestation of epistemic vice seems to be excused, but that does not appear to be the case here. I now wish to turn to delusions.

## 4.3 Background: Delusions

In this section I want to introduce delusions by way of a working definition, and by examples, many of which I will return to as we go along. By way of a definition, the DSM–V says (APA 2013):

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence. Their content may include a variety of themes (e.g. persecutory, referential, somatic, religious, grandiose) [...] Delusions are deemed bizarre if they are clearly implausible and not understandable to same-culture peers and do not derive from ordinary life experience [...] The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity.

Many parts of this definition are controversial, and it is substantially different from the DSM-IV version.[20] There is plenty to say about the definition and its relation to earlier ones but for now it suffices to note that the focus in the updated definition has shifted to what we might call the epistemic features of delusions themselves. These features (fixity, degree of felt conviction, persistence in the face of clear contradictory evidence, etc.) are those that have most puzzled philosophers.

---

A bat and a ball cost $1.10.
The bat costs one dollar more than the ball.
How much does the ball cost?

Without the exercise of vigilance, one is saddled with a false belief. This is not an example of self-deception, but it is a nice illustration of how some epistemic vices are enabled by unwillingness. In cases where this is true, there is a foothold for various forms of assessment, including aretaic assessment. (I leave open the question of whether there are cases of 'pure' cognitive bias and what kind of assessment, if any, would be appropriate there.)

[20] In particular, the requirement that the belief be false, that it be based on 'incorrect inference', and that it be bizarre, have all been weakened or dropped relative to the definition in the DSM-IV. These ought to strike us as welcome changes.

To get an idea for the variety of possible contents for delusions, here are some examples of (types of) delusions, individuated by their content:

1. **Delusions of persecution:** Most common content for delusion. Sometimes called paranoid ideation. The subject believes that his or her life is being interfered with from outside (almost but not always harmfully). Occurs in schizophrenia, affective psychosis, and in organic states.

2. **Capgras delusion:** Subject believes that a close friend or family member has been replaced by an impostor. Best known candidate for 'two-factor' account (Davies et al. 2001.) of delusion formation and maintenance, which I will return to below.

3. **Anosognosia:** The denial of illness. Often follows stroke or brain injury and involves denial of following disability, e.g., paralysis. Ramachandran's (1996) patient F.D. suffered a right hemisphere stroke causing left hemiplegia. But F.D. claimed she could walk and clap. Can also occur in schizophrenia, leading patients to refuse to take medication.

4. **Reverse Othello delusion:** Subject believes in the fidelity of his or her romantic partner in the face of strong evidence to the contrary. Peter Butler (2000) reports the case of B. X., who suffered a severe head injury in a high-speed car accident. Despite the absence of contact with his romantic partner, he subsequently 'developed an intense delusional belief that [she] remained sexually faithful and continued as his lover and life partner' (Butler 2000, 86). I will discuss B. X.'s case extensively below.

I should note just in passing that, despite the language in the DSM (and the language I have used here), it is a matter of some dispute amongst philosophers whether delusions should count as doxastic states. However, in what follows I will be assuming that delusions are best thought of as beliefs.[21]

[21] One reason is simple: I am very sympathetic to the idea that belief is at bottom a concept we use to understand other agents, to explain and predict their behaviour in terms that are readily understandable to us, and generally to calibrate them in the space of reasons. To say that someone believes $p$ might mean things are various as (i) they are inclined to act on $p$; (ii) they are inclined to report $p$ in speech; (iii) they are inclined to use $p$ as a fixed point in practical or theoretical reasoning; (iv) they have a certain felt conviction in the truth of $p$; (v) they treat the question of whether $p$ to be largely settled etc. (Scanlon 1998). So, on this way of thinking about it, belief is a kind of syndrome with no essential features, and someone can be thought to count as believing that $p$ by exhibiting some number of the marks of beliefs. My sympathy with the idea that delusions ought to count as beliefs stems largely from the incontrovertible way in which delusional subjects satisfy, albeit in shifting and sometimes patchy ways, these criteria. In particular, it is very difficult to deny that patients with delusions *take themselves* to believe the things in question. They are subjectively experienced as ordinary beliefs and indeed, one's degree of felt conviction in a delusion can often greatly exceed the conviction one might experience in ordinary belief. As Sims puts it (Sims 2003, 141–142):

> 'It cannot be stressed too often that patients believe their delusions literally: subjectively,

## 4.4 Responsibility and Delusion

Now that we have a working understanding of both self-deception and delusion on the table, and a sense how self-deceivers are typically responsible for their self-deception on my view, our question becomes: are some delusional subjects self-deceived? Does self-deception play a role in forming and maintaining at least some delusional beliefs? Here I will argue that we should say 'yes'. This may seem surprising not least of all because self-deception typically concerns matters which are much more 'garden variety' than the bizarre contents of delusional belief, however, not all delusions have such bizarre content, as we shall see. And even where the content is bizarre, there is room for motivation to be playing a role that might imply self-deception is at work.

If my account of self-deception is correct, it seems to provide relatively straightforward criteria for assessing whether self-deception is implicated in delusional belief. We must only ask whether it is true that the agent has failed to confront, for motivationally biased reasons, manifestly available evidence that would overturn her belief. What remains, however, are two tasks which are not so straightforward: first, we must try to determine whether any actual delusions satisfy those criteria, and further, we must determine what kind of responsibility, if any, that would ground. Let us first address head-on the question of whether any delusions can be thought to fit my model of self-deception. The most plausible candidate for such a case is the Reverse Othello delusion.[22]

---

> delusions are completely different from fantasy. Patients do not describe them 'as if' they existed. The reality is 'known' with the unconcerned certainty that the undeluded person assumes for the concrete events and ideas of his own life, such as the floor being solid...[A] man who believed that American battleships were sailing down the main street of Birmingham UK (100 miles from the sea), had the refined social conscience to report this to the police!'

This example is a nice illustration of how delusions may exhibit some of the marks of belief with clarity and sharpness, even while exhibiting many of the negative epistemic features which are characteristic of them. This subject is using his belief that there are battleships sailing down the main street of Birmingham as the basis for speech, inference, and indeed concern (i, ii, iii) *because* of his degree of felt conviction (iv) in it. Moreover, it seems that the very fact that he takes such a thing to be a reason for concern shows that his belief exhibits a degree of coherence with his other beliefs, beliefs, e.g., about geopolitics and nationhood (not to mention a whole lot of beliefs about military hardware, the nature of peacetime etc.) which together suggest that what is happening is cause for some alarm. The belief is no doubt implausible, and we can imagine that it exhibits a high degree of fixity and resistance to counterevidence, but that should not disqualify it from counting as belief.

[22]The Reverse Othello delusion is noteworthy among delusions for not having the same kind of bizarre content that most delusions have. It may seem in this respect seem tailor-made for someone who wants to defend the claim that there is overlap between self-deception and delusion. A critic might say: 'Most delusions involve believing a highly bizarre content, and it is plausible that (part of) what is distinctive about being in that state is *how* the subject comes to have that attitude towards that content (perhaps, e.g., it is caused by unusual perceptual experience.) But then there will be a gap between self-deception and delusion, one that is missed if we focus only of belief maintenance'. My response is twofold: First, trying to figure out how delusions are formed is surely to be counted as one of the chief aims of the neuropsychology of delusions. And it is very plausible that for a great many of them there will be abnormal mechanisms at work that partially explain (among other things, perhaps) the bizarreness of the delusional content. And this may mean that

## Reverse Othello Delusion

Recall Peter Butler's patient from before, B.X.. B.X. suffered severe head injuries in a car accident. As a result of the crash he was left quadriplegic and unable to speak without the use of an electronic communicator. According to Butler, in the initial stages of his illness he expressed both insight and 'intense emotional response to a massive disability and a fracturing of his interpersonal relationships' (2000, 87). However, in the year following his injury, B.X. gradually developed the delusional belief that he was still in a successful romantic relationship with his former partner (who left him following his injuries) and even claimed that they had recently married, occasionally claiming that he needed to leave treatment to return home to his wife.

B.X.'s appreciation of his injuries is important. He is trying to come to terms with the significance of an irreversible life-changing calamity, and seems to be doing it head-on. But there is a limit to how much such change he can accept at once without falling apart. Butler characterizes B.X.'s delusion as protecting him from falling into severe depression, or as we might say, existential collapse. For him, the ability *to go on* is contingent on his believing that his former partner remains faithful to him. To lose her, on top of all of that has already happened, would be, in some sense approaching the literal, unbearable.

In this context it is also important to note that B.X. eventually manages to recover from his delusion. Even when the delusion was at its most elaborated B.X. did not experience any other psychotic symptoms. The delusional belief seemed to dawn on him somewhat gradually, and eventually reached its most elaborated form in the idea that he and his former partner had been recently married. But the delusion also gradually receded, and he came to accept that she had no intention of returning to him. It is as though the delusion held at bay the need to face something that B.X. was not capable of accepting, until such time as he was more fit to do so.

Together these two things suggest that B.X. is reasons responsive generally. His initial sensitivity and insight into his condition are not things that he could have displayed if he had crossed that strange boundary that leads outside the space of reasons altogether. And the fact that he was able to recover more or less on his own suggests that his capacity to be sensitive to epistemic reasons remained intact — for what else other than that very capacity

there is, in some respect, a gap between (some) delusions and (some) self-deception. But the size of this gap, and its significance, will depend, in part, on what the neuropsychological abnormalities in question turn out to be like. We can see this by considering how, when we think of the neuropsychological abnormalities that we *do* know about already, it is still just a good question the extent to which this shows there to be 'gap' of relevance to thinks like, say, our responsibility judgements. When this-or-that neuropsychological abnormality is discovered, it will still be a good question the extent to which that abnormality presents or underlies a philosophically interesting discontinuity with ordinary cognition, agency, autonomy, or whatever.

Second, even if it turns out that the relevant abnormalities do underlie significant discontinuities between the delusional and the non-delusional with respect to belief formation, it may remain true that there is an interesting overlap between self-deception and delusion precisely because the mechanisms of belief formation are not the only ones we must look to if we want a comprehensive picture of the ways delusional subjects are and are not like 'ordinary' subjects.

could he have used to get himself out? — even if it suffered partial muting and redirection.

On Butler's way of thinking of things — to which I am obviously very sympathetic — B.X.'s belief served a protective or defensive function. How did it serve that function? It is plausible that B.X. needed (in some appropriate sense) to believe something which would forge a strong sense of coherence and connection with his pre-injury self. As Butler suggests, the primary challenge for B.X. during his recovery was coming to terms with how dramatically and irrevesibly changed his life had become. Believing that his partner was there, that a dear corner of his otherwise unrecognizably marred life remained as before, could plausibly offer him something to hang on to, some piece of his past life to use as a flotation device while he tries to get himself to shore.

Now, according to my view of self-deception, it seems that B.X. counts as self-deceived. The belief that his partner had not left him made its appearance sometime after the period of insight that Butler describes. This suggests that B.X. had the belief that his partner had indeed left him at some point prior to the onset of the delusion. Now let us suppose that as the significance of how his life has been transformed dawns on him bit-by-bit B.X. develops a need to believe that his partner had not left him. In a number of ways such a belief is a good candidate for a life-preserver-belief because it concerns a matter which is indeed of great personal significance for him but, compared to the other things of great personal significance to him which are manifestly in shambles it less often and less flagrantly bumps up against evidence to its contrary which would need to be ignored in order for the belief to persist.

It seems that if B.X. is to persist in his life-preserver belief he will need to avoid confronting evidence which points to its falsity. His case is an interesting one because presumably the evidence which is available concerning the falsity of that belief is given by his memory and the memories of his caretakers. Compared to the body of evidence that would have to be ignored if he were to, say, try to deny his injuries (as some delusional patients do), this body of evidence is quite sparse. A little bit of motivated failure to consult that part of one's memory might be all that would be required. If this is what happened, then B.X. counts as self-deceived according to my view. His belief, however it was formed precisely, is false, and manifestly so. But he manages to persist in believing for a time (as long as he needed to, it seems) and this seems to require making himself somehow impervious to the evidence which he had previously appreciated.

The possible complicity of his caretakers in facilitating his failure to confront or appreciate evidence against his delusional belief is another interesting feature of B.X.'s case. It is also quite readily understandable. Clinicians often have to face the difficult question of whether it is appropriate to confront a subject about their delusional belief and many factors might go into determining the appropriate course of action. Plausibly, it would have seemed to many clinicians that the right course of action in B.X.'s case would be to allow him, as he seemed himself to want to do, to take things one at a time, so to speak. If the self-deceiver is encouraged or facilitated in their motivated omission, I am inclined to think that it may partially mitigate his degree of blameworthiness.

Although I do think B.X. is self-deceived on my view, I do not think he is blameworthy. A very small part of the reason for this might be the facilitation of his caretakers. But far

more important, it seems to me, was the function that the self-deceptive/delusional belief was playing for B.X. at the time. If it really was (perhaps the only thing) keeping him together, then I think we are right to see it as an excusable trade-off between negative epistemic value and significant improvement to overall well-being. This doesn't change the facts concerning *whether* B.X. in fact deceived himself, but it is certainly relevant for determining what the appropriate attitude is to take towards him in the light of his self-deception. I chose to express this by saying that the self-deceptive omission, and the resulting delusional belief are attributable to B.X., but that he is not blameworthy for the subsequently persistent belief because he has a weak excuse.

What of the aretaic assessment that is typically grounded by attributability-responsibility? Is B.X. injudicious in the same way that A is? Does he display a negligent quality of will? He may. The moral hazards of self-deception — risk of harm to self or others, for example — are there just as much in his case as in others. But the weak excuse that is available in B.X.'s does more than just block his blameworthiness. It also provides ground for counterbalancing, or perhaps undermining, the aretaic assessment that would normally apply.[23] Excuses of this kind work as follows. Suppose I am tasked with delivering some valuable cargo. If, on the way down the only available path, I encounter a hairy spider and decide to turn back, risking the cargo in the process, I am pretty clearly guilty of cowardice. The action of fleeing is attributable to me (and is the grounds for finding me cowardly), and so too am I blameworthy (liable) if the cargo is lost. On the other hand, if I turn back risking the cargo because there is a grizzly bear on the path, I do not display cowardice, but perhaps prudence. (Or, if one prefers, I display cowardice tempered with prudence.) For the same reason that the negative aretaic assessment of me would seem inappropriate, I submit that I am not blameworthy should the cargo end up lost; the very thing that excuses me from blameworthiness also undermines or counterbalances the judgement of cowardice. This seems to be what is happening in B.X.'s case. His flight from the truth is analogous to my flight from the bear: it comes with risks that we all recognize, but it is not undertaken lightly or negligently. The quality of will that he displays, against the situation in which he finds himself at the same time makes blaming him, and finding aretaic fault, inappropriate.

It is worth mentioning that I want to resist saying that B.X. has a strong excuse for his self-deception. To say this would be to deny B.X. the appropriate ownership over the strategy that he deployed for getting through. I have spoken of his psychological *need* to believe as he did, but I did not mean to suggest that his deceiving himself was something he was literally compelled to do. And more importantly, merely being compelled to do something in this sense might not be enough to constitute a strong excuse. I said that if you only stepped on my foot as a result of being pushed, you would have a strong excuse because the action would not be yours. What if you also, simultaneously *wanted* to step on my foot? Then your action would have a cause outside of you, but would also be an expression of your will. If we were in this situation, we would have to do hard work to figure out which thing

[23] I make this qualification because I want to remain neutral with respect to whether the virtues are necessarily unified.

should properly be considered the reason for your bodily movement. I don't have a general procedure for coming to answer questions of this kind, but I think it is safe to say here that B.X. is not overdetermined in this way. It is clear that what he does is a manifestation of his will, even if there is a sense in which he must do it, because he wills to do as he must. This would not be true of you if you were both shoved and malevolent; your will was not to be shoved, even though it may have been to step on my foot. Being compelled may not be enough to constitute a strong excuse if one also wills the means.

## Capgras, Two-Factor Theory

While I do think that B.X counts as self-deceived on my view, it is unclear how many other cases of delusional belief will satisfy my account of self-deception. My aim has certainly not been to argue that all delusional subjects are self-deceived, or even that it is the norm. However, I think it is worth pointing out that my approach here dovetails quite nicely with a prominent approach in cognitive neuropsychiatry to the formation and maintenance of delusion which is called the 'two-factor theory', and which I mentioned earlier in connection with Capgras. This raises the possibility that motivational factors akin to those that I think are at work in self-deception may be at work in more cases than is widely recognized. Let me elaborate.

Recall the Capgras delusion. Someone with this delusion believes that a friend or family member has been replaced by an impostor. Understanding of how this delusion is formed was greatly enhanced by the discovery that the human facial recognition system has at least two neurologically independent subparts. The first, which is responsible for 'overt' facial recognition is in the temporal lobe and underlies the ability to explicitly recognize the faces of those one is familiar with. The second, affective, system, which appears to involve the amygdala, produces Skin Conductance Responses (SCRs) — *covert* recognition — when subjects are exposed to faces they are familiar with, even if they fail to recognize the face overtly. This is what is thought to be at work in people who have prosopagnosia.

This insight led cognitive neuropsychiatrists studying Capgras to wonder whether the two facial recognition systems were doubly dissociated – that is, whether each was independent of the other and whether there could be people who had the 'opposite' of prosopagnosia. Such people would overtly recognize familiar faces, but would be left without the typical accompanying affective response. Could this be what was causing Capgras? The patient would see his wife, and would accept that the person before him bore an exact physical resemblance to her, but the experience would be entirely without the ordinary feeling of familiarity. It is, perhaps, only a small leap from there to the idea that this person before me, while she looks exactly like my wife, must be someone else.

The two components of the facial recognition system are now largely thought to be doubly dissociated and the abnormal experience of seeing someone who looks exactly like a loved one, but who feels somehow alien, is thought to be involved in Capgras. There is, however, a problem. Not everyone with damage to the covert facial recognition system develops the delusion. Even though these patients are having the same unusual experience as the Capgras

patients, they don't form the delusion. So, something else must be required to fill in the gap between the unusual experience and the subject eventually forming or endorsing the belief.

The debate over what this 'second factor' could be is ongoing, and there is no consensus. But the idea that some kind of two-factor theory is probably correct, at least for some delusions like Capgras, seems enormously plausible. There must be some role for the abnormal experience to be playing, but if that doesn't take us all of the way there, there must be something else at work.

Many of those who pioneered the two factor theory were responding to an idea, tracing back to the work of Brendan Maher beginning in the '70s (Maher 1974), that delusions were largely rational responses to highly unusual experiences. Maher himself thought of his work as a direct challenge to Jaspers' claim that delusions were un-understandable. If some delusions could be understood as rational responses to a certain special kind of experience, experience with a certain kind of force and character, then the content of the beliefs that those experiences gave rise to could be readily understood.

Whether this rational connection can be maintained, and what it means for self-deception depends on how we think of the relation between the first and the second factor. One way of getting at this is to ask how specific the representational content of the abnormal experience is. For example, according to Coltheart (2005), the Capgras patient does not *experience* that his wife is an impostor; rather, an unconscious system predicts that seeing his wife should be accompanied by a certain autonomic response which fails to occur. He thus forms the Capgras hypothesis as an attempt to *explain* the abnormal experience. According to this 'explanationist' account, the representational content of the experience which prompts the delusion is less rich than the content of the delusion itself. According to the competing 'endorsement' account (Bayne and Pacherie 2004), the representational content of the experience prompting the delusion is as rich as the content of the delusional state itself. On this kind of account, the subject does not reach for the Capgras hypothesis as an explanation of his experience but merely takes what is already presented in experience to be veridical.

Obviously, whether there is room for appeal to motivational factors (and whether such an appeal would make a given case count as self-deception on my view), depends on which of these competing accounts is true. On either account, motivational factors (possibly jointly with neuropsychological factors) could be playing a role in generating the anomalous experience. Since the experience is much thinner on the explanationist model, it might be thought that appeal to motivation would be otiose; still, there could be a role for it to play. If, as some philosophers — and increasingly many psychologists — think, it is possible for a subject's propositional attitudes to *cognitively penetrate* her experience, then two subjects may have different experiences even if we hold fixed what is perceived, the perceiving conditions, and the state of the relevant sensory organ. If this is right, then the mere fact that one subject desires that $p$ be the case while the other fails to desire it or desires that not-$p$ be the case, might just be the difference maker when it comes to answering the question, 'Why did the subject have the experience that he had?' — even on the explanationist model.[24]

[24]Of course, this could also be the case on the endorsement model. Indeed, I am assuming that the

If motivation is playing a role in the first factor, that will not be enough for the sufficient condition identified by Self-deception as Omission to be satisfied. When we learn that the subject had some distinctive kind of experience, we just haven't learned one way or another whether there has been motivated mismanagement of evidence which sustains an externally defeated belief over time. However, we may nevertheless be able to learn something about the subject that undergoes such an experience which is akin to what we can learn about the self-deceiver. If we are interpreters of someone who has undergone an experience of this kind, and we learn that this is how it has happened, we come to learn something about the kind of cognitive agent that the subject is.

There is also room for motivation to be playing a role in the second factor,[25] and if it is present, it may go some of the way to restoring the kind of understandability that Maher was aiming for. On either the endorsement or the explanationist account of things, if we have gotten this far, the Capgras belief is already in place, either as an explanation for a bizarre experience or as one given rise to by a bizarre experience directly. Once the belief is in place, there is room for Self-deception as Omission to be satisfied. All that would need to be the case would be for there to be a failure of epistemic agency which is partially motivated by a desire for the world to be as the subject already believes it to be.[26] And as strange as it may sound, the operation of the second factor seems more readily understandable when it is cashed out in motivational terms, or indeed in terms of the kind of mental agency that I think is at work in self-deception. The varieties of human motivation are nearly limitless, and I do not know of any clinical examples that bear this out, but it is not difficult to imagine someone facilitating the maintenance of the Capgras belief for motivationally biased reasons.[27] Perhaps the couple has recently had a particularly acrimonious quarrel

connection between what the subject desires and the content of the experience would be easier to see on this model since the content of the experience is identical to the content of the delusional belief. So, whenever it is plausible that the subject could desire that the delusional belief be true, it will be plausible that the subject desire that the content which shows up in his experience be true. (Of course, whether cognitive penetration works this way — whether desiring that $p$, say, probabilifies an experience with the content that $p$ — is an empirical matter.)

[25]Davies (2010) also discusses cases where motivational factors could be playing a role between the first factor and the second.

[26]Mele denies that self-deception obtains when the subject also suffers from a cognitive impairment, on the grounds that the 'causal contribution [of motivation] may be so small' (2006, 123) that it shouldn't count. I don't see why we should say that it doesn't count rather than say that the contribution that it makes is gradable. And even in cases where (as Mele has in mind) the reflection on the available evidence that is prevented by motivational factors wouldn't have caused the subject to revise his beliefs even if it were allowed to occur, it seems to me that there is a characterologically relevant difference between the agent whose reflection is prevented by motivational factors and the agent whose reflection is not, one that we may well register by calling the former self-deceived and not the latter. When we ask whether someone is self-deceived, we aren't *just* asking which factors are *causally* responsible for sustaining his beliefs. We are asking whether he manifests a certain epistemic vice.

[27]I was told anecdotally of a case where a patient had stopped her medication in an attempt to manage her symptoms without the distressing side-effects that the medication caused. After it became clear that her symptoms were unmanageable without assistance her psychiatrist recommended, to her great dismay, that she restart the medication, to which she responded 'You're not Dr. X! He would never treat me this way!'

and it would be somehow easier to not face the genuine article just yet; perhaps he has been secretly yearning for a divorce and this would save him the trouble; perhaps he has a motivation which only years of deep analysis would uncover. Any such motivation, if it were to underlie and facilitate the acceptance of the Capgras hypothesis, would be grounds for thinking that we had a potential case of self-deception here. The availability of an explanation of this kind greatly reduces the sense that the delusion is un-understandable by bringing the psychological dynamics of the subject into the focus of ordinary intentional explanation.

## 4.5 Conclusion: Innocence

I have tried to bring a number of distinctions between types of responsibility to bear on the question with which we began. I have also put the notion of self-deception to work in a way that I hope has been doubly illuminating: since we have defeasible but determinate antecedent judgements about the responsibility-status of self-deceivers, asking whether someone's conduct can be assimilated to a self-deceptive paradigm can help us think about the ways in which they may or may not be responsible. Delusions can also help us understand the ways in which our ordinary notion of self-deception can be extended to include, e.g., cases where the self-deception is attributable but not blameworthy for very good reason. Using self-deception as a tool for thinking about some delusions also forces on us the question of what a subject's *motivations* are and this question can only be answered by (suitably supplemented) ordinary interpersonal interpretation. Motivations can partially constitute nodes of intentional agency and reminding ourselves about the motivations of subjects with delusions and the role that such motivations may play in our assessment of them can serve as a general bulwark against slipping too easily into thinking of them as outside of the scope of ordinary assessment and understanding altogether.

It is telling that when we bring to bear the tools that I have recommended for thinking about responsibility for delusions we find that there is a good case to be made that the subject is excused. I take this to be in keeping with something Lisa Bortolotti has recently argued for (2016, 2015), viz., what she calls the 'epistemic innocence' of some delusions. She says that a delusion is epistemically innocent if it confers significant epistemic benefits which could not be achieved otherwise. Bortolotti is focused on cases where some negative epistemic consequences are embraced for the sake of otherwise unattainable *epistemic* benefits. I agree with Bortolotti that the notion of epistemic innocence is of clinical and conceptual value. What I hope to have done here is to have introduced what might be thought of as an

This suggests that there are cases of patients forming the Capgras delusion *without* any underlying bizarre perceptual experience and where motivational factors seem to be doing most of the work. Conversely, there have been cases reported of people experiencing the Capgras delusion with their pets, or with inanimate objects (Islam et al. 2015), where motivational explanations seem far less plausible. The size of the gap that needs to be closed by the second factor is evidently highly variable, as is the force of the motivational component. But this doesn't show that motivational factors are not playing a role in some cases of Capgras, it just shows that the role played may be greater or lesser (or perhaps zero) depending on the case.

expansion of that notion of innocence to cases where the negative epistemic consequences are traded off against *non*-epistemic gains. In order to address such cases we need conceptual tools developed from the more general standpoint of moral theory. Taking up this standpoint — taking seriously the possibility that assessment might here really be appropriate — has, I hope, revealed a more comprehensive and detailed picture of what that innocence consists in. There is a kind of innocence which may only be possible against a backdrop of possible guilt.

# Chapter 5

# Addiction

*You become a narcotics addict because you do not have strong motivations in any other direction. Junk wins by default.*

William S. Burroughs, *Junky*

*For self-reflexive beings, the ambivalence of addiction is built into its mechanism: It enslaves by appeal rather than by brute force.*

Gary Watson, 'Disordered Appetites'

## 5.1 Introduction

Addiction is obviously an extraordinarily complex phenomenon. There are countless devoted professionals working in specialized fields investigating how to prevent it and how to treat it; trying to find the neurological, psychological, neuropsychological, genetic, historical, cultural, social, and political *causes* of it; quantifying and minimizing the costs that it wreaks on our health and human services systems, and so on. This chapter will have a comparatively narrow focus. It will be a philosophical investigation into the extent to which addiction impairs morally responsible agency. I will be working within the reasons-responsiveness theory from Chapter 2, and will make use of the framework of excuses from Chapter 4. I will also appeal to a body of recent empirical literature on 'willpower' that I take to be

relevant. My thesis will be that there are two ways in which addiction might impair an addict's reasons-responsiveness. I will argue that the first way, which I will call a failure of 'recognition', may not constitute an excuse if self-deception is involved (as I argue it may be.) However, I will also argue that the second way, which I will call a failure of 'reactivity', does constitute a gradable excuse. These two impairments correspond to the two distinct sets of reasons-responsiveness capacities that morally responsible agents have. First, I will take some time to distinguish between those two capacities, relate them to addiction, and use that as a segue into a very broad-stroke introduction to how I am thinking about the phenomenon of addiction. I will then situate the thesis that I will argue for inside a broader dialectic concerning the extent to which addicts are morally responsible agents. I view both of the two positions typically taken in this dialectic to be ultimately untenable and situate my own position as intermediate between them. One of the positions I wish to reject is supported by the so-called 'Brain Disease Theory' of addiction and I devote a section to critiquing that theory. This is followed by three sections in service of positive argument for my thesis. The first of these is an argument that we ought to countenance *intentions* alongside beliefs and desires in our basic moral psychology. I then discuss each of the two failures of reasons-responsiveness in more detail.

## Two Capacities

The two ways that addiction can impair an addict's reasons-responsiveness capacities correspond to two distinct but related capacities that an agent must have if she is to be fully reasons-responsive. The first capacity, which I will call *recognition* is the ability to appreciate that such-and-such *is a reason* which bears on some practical or theoretical question and that such a reason has a more-or-less determinate degree of force relative to other (potentially countervailing) reasons. To have this capacity is to be able to see things *as reasons* and to therefore to see that they count in favour (or against) a certain course of action or the truth of a certain proposition with some degree of force. This appreciation need not be, it will be recalled from Chapter 2, deliberate or conscious. There are ways of being sensitive to reasons, of allowing reasons to play a role in the determination of what one does or believes which fall short of explicit, conscious understanding *that* such-and-such is a reason with such-and-such degree of force.

The second capacity, which I will call *reactivity*[1] is the capacity to act on the reasons that one has appreciated.[2] Obviously, there is no hard and fast line which separates our powers of reactivity from our powers of recognition, and this is especially so on the *de re*

[1]The two capacities that I have distinguished roughly correspond (*modulo* the proviso about *de re* recognition) to the capacities distinguished by Fischer and Ravizza (Fischer and Ravizza 1998) under the headings 'responsiveness' and 'reactivity'. I prefer not to call the first capacity 'responsiveness' and instead to think of both capacities as required for the theory-eponymous reasons-*responsiveness*.

[2]I have been talking about epistemic reasons and practical reasons together up to this point, but whether there is a such a distinct capacity with respect to epistemic reasons is, in large part, the bone of contention between doxastic voluntarists and non-voluntarists.

responsiveness view: often acting in a reasons-sensitive way just is a way (even perhaps the most appropriate way) to recognize the force of a reason. Examples where we think the situation calls for swift action illustrate this phenomenon well, such as when one sees that another requires immediate aid and rushes to provide it.

Nevertheless, we are all too familiar with cases where the appreciation of the force of a reason leaves us cold or otherwise unable to translate that appreciation into action. Standards cases of weakness of will are of this kind. If I am weak of will my all-things-considered judgement is that $\phi$-ing is the thing to do, yet I fail to $\phi$. This is clearly a failure of reactivity that *presupposes* a recognitional success. This raises a question: given the way I have distinguished between recognition and reactivity, are *all failures* of reactivity against a background of recongitional success going to turn out to be cases of weakness of will? After all, if the judgement that $\phi$-ing is the thing to do is anything *less* than all-things-considered, it won't be a failure of reactivity should I fail to $\phi$. But, then, if the judgement that $\phi$-ing is the thing to do *is* all-things-considered, and I fail to $\phi$ — that is, I exhibit a failure of reactivity — don't we have a case of weakness of will?

My answer to this is to say that it is almost correct, and further to insist that it wouldn't be so bad if it were. However, it is only *almost correct*. I wish to deny, as I did in Chapter 2, that the only form of reasons-recognition is conscious judgement. Once we acknowledge the various forms of *de re* recognition, there is room for cases where an agent, like Huck Finn, *recognizes* some reasons although he fails to make a judgement about them. This then opens up the possibility that he could fail to act in accordance with those reasons. This would be a case interestingly like weakness of will, but it would not involve the kind of clear-eyed motivational conflict that weakness of will is standardly thought to involve. The Huck Finn case, as it is standardly discussed, is referred to as a case of 'inverse akrasia' to emphasize that Huck, by acting against his better judgement, is able to do what is best. However, let us imagine the case as proceeding just as normal right up to the moment of action, except that when the time comes Huck fails to help Jim. By hypothesis, if the *de re* responsiveness understanding of the case is correct, Huck had come to recognize some reasons — to wit, those provided by Jim's humanity — to help Jim escape. His sensitivity to those reasons is what we would appeal to to explain Huck's helping Jim escape if indeed he does help Jim. But the fact that he doesn't help Jim escape doesn't show *by itself* that Huck did not *de re* recognize the reasons provided by Jim's humanity. It just shows that those reasons were not operative (or not decisively operative) in Huck at the moment of action. If we accept that such a case is possible, we will have a failure of reactivity which is not, technically speaking, a case of weakness of will.[3]

[3]It might be objected that there could be no case of mere *de re* recognition of reasons that was not accompanied by action on the very reasons thought to be recognized on the grounds that nothing short of reactivity would show that the agent displayed the relevant kind of recognition in the first place. A full answer to this worry is beyond the scope of the current discussion and will require a foray deep into the metaphysics of the propositional attitudes. However, the imagined objection relies on the assumption that nothing short of overt behaviour motivated by recognition of reasons could provide good grounds for positing that the agent was in the psychological states that facilitate *de re* recognition of those reasons.

In any case, setting aside the difficulties with distinguishing the capacity for recognition and the capacity for reactivity in a theoretically perspicuous way, it is clear enough how impairment of the two set of capacities will present volitional obstacles of different kinds. Concerning addiction, this perhaps corresponds to a familiar bit of twelve-step ideology: admitting she has a problem does mark a real transformation in an addict's process toward recovery. This is because in doing so the addict has achieved a modicum of self-understanding that very well could have continued to evade her for years to come. Indeed, not all addicts ever do achieve it. In so doing, she has overcome one obstacle to the effective exercise of her agency, only to encounter a new one. Sticking with one's resolution to quit is one thing, and presents its own challenges, but you can't even get started in this connection until you *make* the resolution to do so, which, it would seem, requires the ability to appreciate the reasons for doing so. The difference between these two abilities is important for a philosophical understanding of addiction because they correspond to two different types of agential impairments that addicts face. The first is faced by someone who doesn't yet recognize (or won't admit to himself) that he has a problem. It can take an awful lot to get an addict to recognize his own plight, sometimes requiring him to, as they say, 'hit bottom' in order to appreciate the situation.

But of course once an addict has recognized that she has a problem there is still a significant kind of obstacle she will have to face. Those on the outside may be forced to watch helplessly, and are tempted time and time again towards intervention, as the addict promises and reneges, avows and capitulates, until eventually we are forced to the conclusion that she has lost the ability to bring her desires and her behaviour into alignment with her considered evaluative judgements. It is not that she has lost the ability to understand and correctly judge about what things are on the whole good for her, or what she should, all things considered, *do*. Rather it's that these judgements have somehow become uncoupled from what she in fact does.

The difference between failures of recognition and failures of reactivity is a logical one: one cannot take as a sound basis for action — and *ipso facto*, cannot act on the basis of — that which one is not even able to apprehend and appreciate. And it may just be that admitting you have a problem is enough to get you in a position to appreciate what you formerly could not. And this can mark a real difference. For someone who has crossed this threshold, we generally assume that his capacity for making judgements about and commitments concerning what he ought to do is largely intact. Indeed, we assume it is that very capacity which finds expression, however impotent, in the familiar litany of promises and self-remonstrance. To be sure, someone who has made it this far still has a long way to go, but there is all the difference in the world between having the ability to appreciate reasons, however impotently, and not having that ability at all.

---

This assumption further relies on the assumption that what propositional states someone is in is primarily a matter of what state they can rationally interpreted to be in. I am very sympathetic to some version of the latter assumption. That is, I am sympathetic to the so-called 'interpretationism' (see §2.3) of Dennett (1981) and Davidson (2004c), but I am not sympathetic to the first assumption, which is a specification of it.

These impairments are all the more striking because there is another way in which the addict's capacity for making self-determining choices remains largely intact. I mean this in two senses: First, the impairment to the addict's agency is, both in terms of recognition, and in terms of reactivity, quite circumscribed. I will return to this point below, but for now it will suffice to point out that someone with an addiction to a particular substance, say heroin or nicotine, may find that they have no difficulty forming judgements and commitments, and then executing actions in accordance with those judgements and commitments, concerning, e.g., whether stop signs should be heeded or whether one should keep promises. To be sure, the effect of the addiction may have a considerably long reach into an agent's practical decision making network, as more and more things become related to, or subservient to, the end of procuring and taking the drug. But this is a development of the phenomenon and not, it seems to me, essential to it, and in any case, it would be extreme to suppose that such a development could, even in theory, come to constitute a global impairment to agency. The second sense in which the addict's agency seems somehow to remain intact is that each of the individual episodes of addictive behaviour — drug-takings, say — seem themselves to be exercises of intentional agency. It is of course true, and we shall return to this later as well, that part of what is distinctive of the addict's predicament is the strength of the desires he has to successfully overcome if he is to resist taking the drug again. This is pretty clearly a significant motivational obstacle. But we should resist thinking of the addictive impulses as literally irresistible forces. Although this may seem like a theoretically tempting move to make — it wouldn't require postulating any global breakdown of agency and yet would seem to get us quite a way towards explaining the sometimes extraordinarily destructive power of addiction — there are many reasons for not going that way. For one thing, it would come at the cost of making it utterly unintelligible how anyone could successfully resist those impulses and achieve recovery without literal physical manipulation from without. Second, this position is empirically untenable (as I will discuss below). It also threatens to lead us into a confused conception of how desires relate to the will, as I hope we shall see.

What we have, then, is a picture of addiction as a potentially strongly debilitating condition, involving an impairment to agency of variable severity and locality, which nevertheless operates through the motivational and deliberative psychology of the subject. Addiction would not be puzzling if it did't run through our psychology in this way. There would simply be no puzzle if we were creatures entirely determined from the outside, with no capacity for self-control. All it would take to trap a creature such as that in a pattern of self-destructive behaviour would be a set of sufficiently strong forces moving him in that direction. It will be a large part of my aim here to attempt to resist thinking of ourselves in these terms, as simply subject to forces that move us, without denying that there are significant 'external' forces at work in producing and sustaining addictive behaviour. As an illustration of how I am thinking about this, consider a study published in *Nature* in 1998 (Pianezza, Sellers, and Tyndale 1998).[4] The focus of the study is what Michael Pianezza

[4]This study is representative of many which shows a connection between certain genetic factors and tendency to engage in addictive behaviour which runs through the subject's motivation psychology. I have

and his colleagues at the University of Toronto call a 'genetically variable enzyme', CYP2A6, which is responsible for the metabolism of nicotine into cotinine (this accounts for approximately 60-80% of overall nicotine metabolism). The hypothesis of the study was that there would be 'under-representation of individuals with impaired nicotine metabolism (carriers of the null CYP2AC alleles) in a tobacco dependent population' — that is, if someone doesn't produce the enzyme which turns nicotine into cotinine, it is less likely that they will be addicted to nicotine. And the study bears this out. But it is eminently plausible that this effect is mediated by the psychological effects of having or not having impaired nicotine metabolism. If, as seems plausible, someone with impaired metabolism will simply be *less likely* to experience ingesting nicotine as pleasurable than (or as pleasurable as) someone who does not have impaired nicotine metabolism, the two subjects may be in very different motivational and volitional situations. It will be the goal of this chapter to come to a better understanding of the motivational situation faced by addicts and I believe that appeal to empirical study can help up greatly in this regard.

## A Cartoonish Opposition

Before proceeding further it will be useful to sketch the framework to which pretty much all public debate on the nature of addiction is forced to conform. It is an admittedly cartoonish picture, and I take both positions to be untenable, but it is worth saying why they are so. It is a little bit challenging to say just what the two positions in question amount to exactly, since on reflection they can appear to be little more than flat-footed expressions of a deep disagreement over the basic moral issue at the center of all this: Are addicts morally responsible for their behaviour? Those on what, for lack of a better term, I will call the 'conservative' side are committed to delivering a resounding 'yes'; those on the 'liberal' side, 'no'. But the views are rarely expressed as simply as that. They almost always take on board further ideological baggage. Conservatives say: 'Addicts are responsible because of moral weakness'; 'Addicts know what they are doing is bad for themselves and bad for their families, but they do it anyway. That's a choice. They *deserve punishment* for this choice.'; 'You could always get yourself out of it if you just tried harder'; 'This is the result of too much hedonistic partying'. Liberals, on the other hand, say: 'Addicts have no choice in the matter, they have a *disease*'; 'Addicts' brains are hijacked by drugs, they are literally powerless'; 'Addiction is like diabetes or schizophrenia, you wouldn't put someone in jail for having a disease, would you?'[5] There are obviously a few unhelpful things going on here: the unthinking assimilation of addiction to *akrasia*; appeal to the problematic concept of a disease; the attribution to drugs of the almost magical power to obliterate someone's capacity

chosen this one for its brevity and relative accessibility.

[5]The two positions are actually significantly more confused than that. The conservative position almost always comes with views about the wrongness of taking drugs and insistences that the only effective way to control drug taking behaviour is strong criminal deterrence. The liberal view often comes with (sensible seeming) disapproval of prohibition on both moral and policy grounds and (less sensible seeming) advocation of twelve-step programs and the like as the best avenues of treatment.

for self-control; and confusions between issues of responsibility and issues of criminalization and punishment besides. However, the most obvious and most serious problem here, which we see on both sides, is the mistake of assuming that the answer to the central question ('Are addicts morally responsible for their behaviour') must either be an unqualified 'yes' or an unqualified 'no'.

My objective here will be to find an appropriate middle ground, by appeal to a suitable conception of the *will*. My strategy will be to begin with what is perhaps the most flat-footed basis for the liberal position available, which is often called the 'Brain Disease Theory' (BDT) of addiction. I will argue that BDT is not able to provide support for the liberal position without taking on board unacceptable extraneous commitments. I will then address a more straightforwardly philosophical consideration that could provide a basis for the liberal view, which I shall call 'Humeanism'. I will argue that Humeanism is unable to capture some of the basic volitional phenomena central to our lives as reflective, self-determining agents, and I will attempt to buttress objection into a conception of the will that I hope will help us both to understand what is distinctive about the predicament that addicts find themselves in, and to begin to articulate an acceptable middle-ground position in the disagreement over responsibility.

## 5.2 Against the Brain Disease Theory

What I shall call the 'Brain Disease Theory of Addiction' (BDT), as I understand it, is committed to the claim that there is such a state as the state of being addicted, and that that state is identical to a complex state of the addict's brain, a state which is caused by repeated exposure to addictive substances or repeated execution of potentially addictive behaviours. The nature of this state can then be specified with various levels of precision. But BDT is also usually understood to be bound up with a couple of 'extra-theoretical' considerations: promoting effective treatment, promoting an increased sense of empathy for those suffering from addiction, and (or perhaps, *by*) eliminating the stigma attached to addiction caused by the thought that addicts are somehow responsible for their condition. Since this perspective on addiction has been developed by both treatment professionals and neuroscientific researchers, the emphasis may shift, depending on who you ask: a clinician may emphasize the efficacy of treatments which intervene at the neurochemical level, and infer from this that the condition is one of the brain, one which perhaps undermines responsibility somehow; or, convinced by neuro-imaging studies that the condition is essentially one of the brain, one that bypasses the will altogether, a neuroscientist or psychiatrist may recommend on that basis some treatments over others and agitate for increasing empathy and reducing stigma. Nora Volkow, director of the National Institute on Drug Abuse, — a federal agency with a current annual budget of more than a billion US dollars — articulates the core of the theory and the desired 'extra-theoretical' results as follows (Volkow 2015):

> We in psychiatry [should] embrace addiction as a chronic disease of the brain,

> where the pathology is the disruption of the areas of the circuits that enables us to exert free will, that enables us to exert free determinations. Drugs disrupt these circuits. The person that is addicted does not choose to be addicted; *it's not a choice to take the drug.* Many times they take it and they'll say it's not even pleasurable. "I just cannot control it" Or they'll say, "I have to take the drug because the distress of not having the drug is so difficult to bear".
>
> If we embrace the concept of addiction as a chronic disease where drugs have disrupted the most fundamental circuits, that enable us to do something that we take for granted — make a decision and follow it through — we will be able to decrease the stigma, not just in the lay public, but in the health care system, among providers and insurers. So that patients with mental illness do not have to go through obstacles to obtain the evidence-based treatments, so that they don't have to feel that shame, they don't have to feel inferior, and perhaps we will be able to feel empathy for a patient suffering from a disease we call addiction.

It is the relation between, as it were, the core of the BDT, and these 'extra-theoretical' results that I am interested in, specifically the claims it wants to secure about responsibility. It is of course desirable to promote empathy, and to increase the efficacy of various treatment methods, but it is easy to see that the relation between the core of the theory and what it is supposed to imply about responsibility is more fundamental than any of these other results. It is *because* the theory somehow shows us that addicts are not responsible (in the relevant sense, for the relevant things) that they should be treated with more empathy, and *because* we understand that addicts suffer from a condition of the brain implying they are not responsible that we are able to intervene with treatments at the correct level in order to maximize efficacy. The present question is how to understand the relation between the BDT and the claim that addicts are not responsible agents.[6] How might we bridge this gap?

## Cause and Control

Since for every psychological state there is a brain state, one might wonder whether the BDT adds anything we are not already committed to if we are physicalists. In addition to just stating what seems like a pretty straightforward consequence of physicalism, BDT also says something about the causal history of the state in question. The thought seems to be: science has revealed to us that certain kinds of neurochemically abnormal states of the brain induced by ingesting intoxicating substances or performing other rewarding behaviours can cause long-lasting damage in the brain's reward system. It is not surprising, perhaps, that there are neural correlates of both (say) the experience of taking, and the decision to take, a drug, on the one hand, and being such as to have acquired a very strong appetite

[6] There may be several ways for someone's responsibility to be undermined, but Volkow puts it in terms of the addict having no *choice*, that he has *lost control*. I may thus talk somewhat interchangeably about undermined responsibility and loss of control when the distinction does not appear relevant.

for the drug, on the other. What is perhaps revelatory is that the whole story can be told without making any reference to the agent's character, beliefs, desires, strength of will, or social environment; if your brain is flooded with dopamine enough times, it'll switch to operating in an abnormal pattern which underlies and explains the addict's peculiar and compulsive-seeming behaviour.

Suppose we accept this. Does this get us any closer to any claims about addicts' responsibility? Volkow seems to be asserting that the understanding of addiction that brain science has afforded us shows that addicts lose their capacity for self-control: they cannot help but take the drug. And indeed, this would seem to have fairly obvious consequences for whether addicts are responsible for taking the drug. But is this implied by BDT, as I have stated it? I think it is pretty clear that it is not. BDT provides a (kind of) analysis of what it is to be addicted, that is BDT identifies the state of being addicted with being in an abnormal, highly complex brain state with a certain causal history, but so far the theory itself has given us no reason to think that addiction involves a fundamental disruption of someone's ability to, as Volkow says, 'exert free will'.

What is obviously needed is a more detailed understanding of the nature of the brain state BDT identifies with being addicted, and the causal history of this state may turn out to be key. Performing an addictive behaviour causes the brain to be in some abnormal state (flooded with dopamine, say) and repeated inducement of that state is understood to result in the state which is characteristic of addiction (lesions in dopamine sensitive areas of the reward system, say). Could the nature of the state, understood in this way, be enough to undermine an addict's responsibility? Perhaps the idea is: Repeated overstimulation of dopamine receptors leads to lower baseline dopamine receptivity in people with a long history of substance abuse.[7] These lower baseline levels produce both tolerance and appetite: the craving for the drug accompanies the understimulated state, and as the 'deficit' of the state deepens, more and more of the drug is required to produce the same effects. Reduced levels of activity in the reward system as a whole also blunt the rewardingness of (and the concomitant motivation to pursue) life's other valuable pursuits strengthening the centrality of the drug in the addict's practical decision-making network.

In order for this to undermine the addict's responsibility, however, it seems that more is still needed. Someone who is skeptical that the BDT can get us to the position that addicts' responsibility is compromised or undermined, may always insist that what has been discovered are merely neural correlates of items in the agent's psychology — addictive impulses, say, *pangs of appetitive desire* — and it is those irreducibly psychological states, and their interaction with the agent's other psychological states that figure in the causal explanation of the agent's action. Whether the agent has control over the states in question, can exercise self-control *against* the power of those states, is moved, or left cold, or overwhelmed by

[7] Although drugs of abuse can be classified into two groups based on which mechanism they exploit to produce this result, they all produce basically the same result. One class of drugs (such as marijuana and opiates) mimic the action of endogenous neurotransmitters; the other class (which includes cocaine) inhibits the normal reuptake of these neurotransmitters. In both cases the brain behaves as if it is (or it is) flooded with abnormally high levels of dopamine.

them on any give occasion, is a further matter. So, we may think we have found the neural correlates of a particularly strong, unruly, addictive appetite, but we may be nonreductive physicalists about that appetite and the other mental states relevant for satisfying or resisting that appetite. That is, the non-reductive physicalist blocks the BDT's attempt to close the gap between the core of the theory, and any undermining of the addict's responsibility by claiming that the causal powers of the mental are not reducible to the causal powers of the brain on which it supervenes.

Finding the neural correlates of addictive appetites doesn't tell us anything about the degree of control that addicts are able to exercise. We have admitted one, perhaps particularly powerful, item into the addict's psychic economy, which may be enough to set him apart from other people who aren't subject to anything quite of that sort.[8] But the causal etiology of any given human action is almost always highly complex, and cannot rightly be said to be determined by a single token of a psychological state. Even in ordinary trivial cases, my $\phi$-ing rather than $\psi$-ing is caused by much more than just a desire to $\phi$ — my beliefs about the costs of $\phi$-ing are relevant, as are my beliefs about the benefits of refraining, my background beliefs, my other desires, the current direction of my attention, my overall level of comfort, my other normative commitments, my habits and dispositions as they relate to $\phi$-type activities, etc. Why should these other states not be relevant for the addict? Further, the capacity for reflective agency which is present in ordinary adult human beings in virtue of which we respond to, weigh (not necessarily consciously or deliberately) and ultimately act on the force of reasons is active in a non-trivial number of cases of human action.[9] Why should we think this capacity is rendered inert or overridden because we have discovered an abnormality in the addict's reward system? To get to the claim that the addict is unable to exert self-control it seems the BDT must claim that being in the relevant brain state causally necessitates the addictive behaviour. It seems that there are two ways to do this: (i) accept in outline the picture of agency that I have just sketched, but claim that the strength of the addictive impulse or appetite is so strong as be literally irresistible; or (ii) reject the picture of agency just sketched by denying that (certain) actions (performed by addicts) are in any meaningful way caused by the agent's mental states.

### Against the first option

It is of course true that addiction involves *some kind* of impairment to one's ability to exercise one's agency. One is tempted to say that this is essential to any proper understanding of the phenomenon. But even for all the obstacles to agency that addiction throws up, these

[8] Or maybe not. Perhaps the difference is just one of degree.

[9] Many actions, of course, occur much more 'automatically' or unreflectively than this. Not only needn't the weighing of reasons be conscious or deliberate, the action may be less of a response to reasons at all and more of a brute habit. Plausibly, the forming of (in particular) an addictive habit will correspond to the changes in the brain which the BDT says precedes arriving at the addicted state. But the fact that this is a habit which can produce action without reflective or deliberate interference or the exercise of one's capacity to respond to reasons doesn't show that that capacity is undermined.

impairments are all the more striking because there is another way in which the addict's capacity for making self-determining choices remains largely intact — and this much the Brain Disease theorist needn't deny, in the addict's case, or in the normal one, on this way of going. Even the most desperate addicts are able to modulate their behaviour when the situation demands it, when for example, a favoured dealer is unavailable, or when an experimenter artificially cranks up the rewards of temporary abstinence (Fingarette 1988, 34–49). We should not overlook the abnormal strength of the addictive impulses, and it is of course true that part of what is distinctive of the addict's predicament is the strength of the desires the he has to successfully overcome if he is to resist taking the drug again. But I think we should resist, if we can, thinking of the addictive impulses as *literally* irresistible forces. I would like to mention a couple of reasons for not going that way.

First, this position has a dubious empirical underpinning. The idea that certain addictive drugs hijack the brain and produce literally irresistible impulses to take them can be traced to a large series of experiments starting in the 1960s on caged animals which purported to show that the animals became utterly compelled to take the drugs to which the experimenters had caused them to become addicted. The late Stanford pharmacologist Avram Goldstein summarizes that research and draws this inference as follows (Goldstein 1997):

> Every addictive drug used by people is also self-administered by rats and monkeys. If we arrange matters so that when an animal presses a lever, it gets a shot of heroin into a vein, that animal will press the lever repeatedly, to the exclusion of other activities (food, sex, etc.); it will become a heroin addict. A rat addicted to heroin is not rebelling against society, is not a victim of socioeconomic circumstances, is not a product of a dysfunctional family, and is not a criminal. The rat's behavior is simply controlled by the action of the heroin.

I take it he means to further infer from this that the behavior of *people* who are addicted to, say, heroin, is also 'controlled by the action of the heroin'. But many things make this a bad inference. For one thing, people aren't rats. But for another, animals kept in cages given the choice to self-administer, say, tap water, or heroin-laced water will *of course* opt for the drug-laced water. Bruce Alexander's groundbreaking and under-appreciated research has done a lot to directly resist inferring what Goldstein does from experiments on caged animals (Alexander, Coambs, and Hathaway 1978, Alexander 2008). Alexander and colleagues showed that animals will forego the the self-administration of drugs if saccharin or social interaction are available as alternatives. Given this, these experiments have not even shown the existence of a literally irresistible impulse in *rats*, rather than, say, inducing the rats to take drugs by the poverty of their condition. That is to say nothing of the difficulty of generalizing from rats or monkeys in a laboratory to human beings in a highly complex social environment. Even if we accepted that the best explanation for the caged animals' behaviours was the operation of an irresistibly strong impulse, the highly complex capacity for normative self-determination possessed by human beings should give us pause before extending such results to human beings.

The second reason why we should resist this strategy is that it would come at the cost of making it utterly unintelligible how anyone could successfully resist those impulses and achieve recovery without literal physical manipulation from without. Some people are not able to wrest themselves from the grip of addiction on their own, but many can, something which cannot be accounted for if addiction produces impulses which are literally irresistible. In fact, recovery, usually unaided, is not just a mere outside possibility for those who suffer from substance abuse or dependence at some point in their lives, it is the norm. The National Institute of Mental Health (NIMH) conducted a landmark study between 1980-1984 to measure the prevalence of psychiatric disorders amongst the U.S. population according to the then-current DSM III criteria. One of the study's most striking findings is that more than half of those who previously met the criteria for drug abuse or dependence reported no symptoms at all by age 24; by age 37, almost 75% are symptom-free.[10] Case studies suggest that this 'maturing out' of addiction is the result of finding meaning in other pursuits, such as career or family, as one grows older.[11] This suggests that addicts are not at the mercy of a behaviour-controlling brain state.

### Against the second option

The second way for the BDT to try to close the gap between the core of the theory and the undermining of addicts' self-control would be to claim that the scientific study of the brain has found the causes of addictive behaviour and that those causes *leave no room* for folk-psychological states in the causal etiology of that behaviour at all. That is, an appeal to could be make to kind of *causal exclusion principle*, which might say something like:[12]

> **Causal Exclusion Principle:** If a higher level state or property M, supervenes on some lower level state or property P, that is causally sufficient for state or property P*, then M cannot cause P*.

On the plausible (perhaps because very weak) assumption that an agent's responsibility for performing an action is *prima facie* undermined if none of that agent's mental states figure in the causal etiology of that action, this may be just the sort of principle that the Brain Disease Theorist will need in order to bridge the gap. But we will have to be careful about which state we take P to be. If, on the one hand, we identify P with the total state of the person's brain which precedes the performance of the addictive behaviour, we will, of course, have identified a state which causally suffices for the performance of that behaviour. And so, if we accept the CEP, along with our plausible assumption about responsibility, we may have to accept that the agent's responsibility is undermined. But, unfortunately, the very same reasoning, if sound, would seem to suffice to undermine the responsibility of

[10] These results are helpfully summarized in by Gene Heyman (Heyman 2009, 70).

[11] Heyman 2009, Ch. 3 contains interviews and case histories for a number of addicts, some recovered. It is a notable theme among those who recovered that they 'found meaning in other things'.

[12] I here assume without argument that the mental supervenes on the physical.

*everyone*. That is, the causal exclusion of the mental in general, when conjoined with the BDT, will suffice to undermine the responsibility of addicts only at the cost of undermining all responsibility.[13] The hope of the BDT was to be able to tell us why *addicts* and not others, have impaired self-control or bear no responsibility, so it will not do to claim that it is because the mental is not causally efficacious in general.[14]

Alternatively, the Brain Disease Theorist could try to identify P just with the unhealthy state of the agent's reward system induced by repeated exposure to an addictive substance. But then, it will be difficult to see why that state should, by itself, causally suffice for P*, performing the addictive behaviour. That state is, we could grant, a relatively stable feature of the agent's brain, but it is also a fairly local one. The agent's brain may be in a whole host of other states, and each of those states may be the subvenient base for any number of other mental states that might be relevant for the performance or non-performance of the action. Why should we think *those* states are causally inefficacious? If we accept what has been claimed so far, the reason we give here should not be the causal inertness of the mental in general. Nor will it do to admit that the mental states supervening on these additional brain states are generally causally efficacious but insist that they are powerless to exert any effect in the addict's case, for that would be tantamount to claiming that the addict is in the grip of a literally irresistible desire and falling back into the first option. Thus, it seems like it might be possible to block this strategy for bridging the gap simply by denying that the antecedent of the CEP principle is satisfied.

In general the debate about mental causation and the CEP has focused on whether there is any way to deny the principle itself.[15] I have sympathy with some attempts to argue against the truth of that principle, especially those that think there has to be a 'proportionality' constraint on causal explanation. In particular, I think there is something to the suspicion that there is somehow a mismatch between non-intentional causal *explanans* and intentional *explananda* — but I have not relied on the falsity of CEP here. Instead, I have been asking whether it can be applied to help the BDT, as I have been saying, 'bridge the gap' between its core metaphysical commitment and desired results about responsibility, and I do not yet see that it can.

## 5.3 Humeanism and the Will

If I am right about the BDT, liberals will have to find support for their thesis elsewhere. The view which I shall call 'Humeanism' or 'The Humean View' may be able to provide such support in a slightly more nuanced form. I will, however, ultimately recommend that we

[13]There is, of course, a kind of compatibilist reply available here, but it would be available in the addict's case as well, so it seems that the issue of compatibilism vs. incompatibilism 'divides through' in the sense that the BDT seems to require the rejection of compatibilism in any case.

[14]This is essentially a version of Kim's (2003) argument that non-reductive physicalism is committed to a kind of epiphenomenalism.

[15]Yablo (1992) has an interesting line on this, as do List and Menzies (2007).

reject the Humean view and I will try to use the shortcomings of that picture to motivate an alternative picture. Since, ultimately, I am trying to make good on the promise to say something illuminating about addiction by way of a conception of the will, I need to be very careful about homing in on the right one. The will has been rung in to do a lot of different work in the history of philosophy. I want to impress, especially to those who are skeptical about whether we need the will and suspect it might be a metaphysical extravagance, that appeal to the will in the history of philosophy has always occurred because it was needed to fill a particular theoretical role. Augustine needed it to secure a non-Manichean theodicy; Descartes needed it as the source of the mental cause of every bodily event. And this, I suppose, is as it should be. I cannot think of a more methodologically respectable principle for accepting entities into our theories — especially, I should say, in philosophical-psychological theories — than a principle which rules in the indispensable (and perhaps we can now say, rules out the extravagant). Indispensability will, of course, be relative to our aims, but my aims are fairly standard. I want a theory of motivation and action which is maximally general, and which doesn't metaphysically overreach, but which, importantly, is not totally flummoxed by the phenomenon of addiction. Let us turn then, to the Humean picture.

## The Humean view

Hume himself speaks of the will as if its exercise accompanies every action, but he is also very clear that it is nothing but an impression (Hume 2000/1738, 2.3.1.2, editor's emphasis):

> I desire it may be observ'd, that by the *will*, I mean nothing but *the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or new perception of our mind.*

So, although the will is present as an impression accompanying all of our actions, it is not itself a separate cause of action. This means that, for Hume, if we want to know what gives rise to action, we can't look at the will itself, but we shall have to look at the 'motivating influences' of the will. In T 2.3.3. Hume characterizes these influences very narrowly by arguing for two claims: (1) 'that reason alone can never be a motive to any actions of the will, and (2) 'that it [reason] can never oppose passion in the direction of the will'. It is worth considering briefly how he argues for these claims to get a better idea of what motivates the picture bears Hume's name.

According to Hume, what we might call mental states come in two fundamentally different varieties. First, there are those that purport, as we might say, to represent the world the way that it is. These belong to the understanding. As Hume puts it: 'The understanding exerts itself after two different ways, as it judges from demonstration or probability; as it regards the abstract relations of our ideas, or those relations of objects, of which experience only gives us information.' The goal of T 2.3.3 is essentially to argue that these states are *not* influencing motives of the will. Those motives are instead members of a totally different class, *passions*: 'A passion is an original existence, or, if you will, modification of

existence and not any representative quality, which renders it a copy of any other existence or modification.'

Contemporary philosophers have developed these remarks into what we might call the Humean Theory of Action.[16] Proponents of this theory typically express it by saying that causes (or explanations) of action are belief-desire pairs, where the desire specifies the end for which the agent acts and provides the motivational force necessary to move the agent to action, and the belief concerns the means to the end specified by the desire. At any time I am subject to a sort of field of forces pointing in different directions, and I represent the world in practical terms with the understanding insofar as I understand this-or-that action to be a means to the satisfaction of this-or-that desire. When I act in the way that I do, I am taking the means to the end which exerts the strongest motivational force on me. The view thus has two components. The first is a negative component which identifies a necessary condition for action. In slogan form: there is no action without desire. The second part is positive: whenever an agent acts, he acts on the strongest of his desires.

Proponents of this theory thus appear to accept in some form Hume's restrictions on what can count as a 'motivating influence of the will'. And this position is both agile and very elegant: the division of labour between intellect and motivation that it relies on can seem intuitively plausible; it captures nicely the idea that there is a limited set of actionable possibilities for an agent at any given time — those actions which in some sense and to a non-zero degree she has desires to perform (and those actions which are suitably related to actions). Perhaps most interestingly for our purposes, the theory can also seem nigh invulnerable to refutation by counter example. Any scenario an opponent can come up with which purports to be an action without the required kind of desire is one that can easily be re-described by a Humean in a way which is consistent with his theory. This is due both to the fact that the concept of desire in play is highly elastic and that it comes to take on a highly theoretical connotation which can make desires seem cheaper than they are in ordinary experience. Meiko, on the business end of the highwayman's .45, gives up her purse. Elijah drinks a can of paint. There must have been *something* which made it, from their perspectives, seem like these things were good ideas which inclined them in that direction or, at the very least, which we can appeal to to make their actions so much as intelligible. Although we wouldn't ordinarily say that Meiko *wanted* to hand over her purse, it was, in her situation, the best option, and she was in a position to see that. We can express this by pointing out that she has a standing desire to preserve her life, and acted so as to see that it was preserved. Elijah may have no reason to drink the paint except that he has a yen to do it. However, with only this in hand we could say that he desired to do it. Of course, the sense in which Elijah desires to act as he acts, and the sense in which Meiko desires to act as she acts, are very different, but if we agree to call that thing, whatever it is, which is present in both cases, a desire, the Humean is happy. We might quite rightly, however, think that

[16]It is more customary to call it 'The Humean Theory of Motivation'. See, e.g., Michael Smith's (1987) 'The Humean Theory of Motivation'. However, I prefer not to characterize it as a theory of motivation, since motivation is one of the *explanans*, not the *explanandum*.

this is highly slippery. Indeed, I suspect that this phenomenon is related to another which is suspicious: *arguments* for the Humean view are in disappointingly short supply.[17]

What I think is going on here is traceable back to the transition from talk of the contents of the understanding and the passions, on the one hand, to talk of beliefs and desires, on the other. Proponents of the theory have elevated the terms 'belief' and 'desire' to the status of synechdoches for the contents of the understanding and the entire class of the passions, respectively, in order to mark the difference between those states which have motivational efficacy and those which do not. I did say — and I do think — that there can seem something plausible in the idea that there is this difference. Reflection on ordinary cases where an agent is not able to get herself to act suggests that we could posit a kind of motivational force that is simply absent. But to do so is to slide into highly theoretical talk of 'desire', which bears very little relation to the ordinary — phenomenological — conception. It will turn out, on this view, almost as a matter of definitional necessity, that any action done for a reason will show the presence of that special kind of very thin mental accompaniment which confers the status of an intentional action onto it. And this claim will require no argument. But it is no longer a psychological — or indeed, an interesting — claim. The no-actions-without-desires thesis seemed to promise to draw an interesting and potentially highly illuminating connection between the psychological conditions of individual agents — understood in familiar and evocative terms: what we are antecedently 'moved' towards — and the limit of their actionable possibilities. Hume's skepticism about the ability of reason alone to move us would indeed have profound consequences for our theory of agency. If it were true, it would seem to show that our capacities for rational self-determination are quite limited. Reason would not be able to break the bounds of practical possibility imposed by a psychological condition that one simply finds oneself in: desires in the phenomenological sense *are* quite impervious to the effects of rational deliberation. But if every action is accompanied by desire in merely the formal — or what we might call 'pro-attitude' sense — and not in the phenomenological sense, we haven't imposed any such independent limitation on that array of possibilities, for it would then become an open question whether the potentially vast variety of pro-attitudes may include states which are responsive to, or wholly dependent on, reflection, deliberation, and volition.[18]

I should be clear: I am not merely complaining that the assimilation of all of the states of the understanding to 'beliefs' and all of the passions — amongst which Hume himself includes (ordinary) desire, fear, joy, grief, fear, hope, aversion (T.2.1.1.4) — to 'desires' is artificial. The plausibility of the theory hangs in interpretive balance with its interestingness: on the reading of 'desire' in the pro-attitude sense, as a merely formal or theoretical component of every action meant to account for the fact that the agent was so much as capable of performing it, and to which the Humean is forced to by even the most elementary of purported counter examples, the theory may well be true. But on this reading it is

[17]Other than Hume's own, which don't really achieve the status of arguments, Micheal Smith's 'direction of fit' argument is the only one that comes to mind, from Smith (1987).

[18]Much of this argument is endebted to Wallace (1990).

likely not to be of much interest, and it would certainly be a mistake to arrive at anything like skepticism about practical reason on this basis. To do that would be to simply beg the question of whether any of the pro-attitudes are judgements or volitions. On the other hand, on the more ordinary, phenomenological, reading of 'desire' the theory offers a highly interesting thesis, which simply appears to be demonstrably false: When I manage to get myself out of bed after a night of restless sleep and facing a looming morning commitment, it may not be due in the least to the operation of any felt inclination do so.

Of course, the Humean will not concede so easily. Presumably the best response would not be to try to defend the phenomenological version of the theory, but rather to try to vindicate the class of motivationally efficacious psychological states picked out by the term 'desire' understood in the pro-attitude sense. In order to do this, it seems, the Humean would have to show that that class is a psychologically natural one of enough independent interest that thinking of all of the varieties of action-impetus that we are familiar with from ordinary experience under the umbrella of 'desire' would be neither artificial nor misleading. Moreover, it would have to be that what unifies this class of psychological states is not only their motivational efficacy, but also that they are simply 'given' to us, in the way that ordinary desires can seem to be. And this seems like a tall order. In particular, it seems that taking seriously the perfectly familiar phenomenon of forming and following through with one's intentions is enough to undermine the prospects of a satisfying reply along these lines.

Before pushing the dialectic with the Humean any further, it is worth pausing to note how a Humean conception of action could provide the basis for a liberal view of addiction and responsibility. According to the liberal, addicts are not responsible for their addictive behaviours, and according to the Humean, the agent always acts on the strongest of her desires. If, further, desires are simply 'given' to us, and are impervious to our reflective judgements, and *nothing else* is a possible cause of action, then perhaps the addict fails to be responsible because she is subject to desires which control her conduct and for the having of which she is not herself responsible. Note that as a possible basis for a liberal view of addiction and responsibility, this strategy might seem to face the same cluster of worries that faced the the BDT. Just as above, we don't want it to be that the addict suffers literally irresistible impulses, nor do we want the addict's failure of self-control to be grounded in a claim of universal failure of self-control. However, for the Humean, the threat of both of these worries, if they are to be taken seriously, comes, so to speak, from the same domain — from within the psychology of the agent — and so there is the possibility that Humean can address them both at a single stroke. What I mean is that, unlike for the BD Theorist, the Humean doesn't threaten to commit us to a version of epipehomenalism — since some psychological states, viz., desires *are* causally efficacious on this view — but it could nevertheless appear to commit us to a kind of control-bypassing causal necessitation *if any desires are irresitible*. Thus, the Humean should simply deny that any desires are like that. But she must do so within the confines of her theory. How is this possible? If no desires are irresistible, what sense is to be made of the claim that the agent always acts on the strongest of his desires? For there to be any space here, it must be possible for the agent to somehow manipulate the effective force of his desires. Thus, with these resources (and only

these resources) the Humean will at this juncture attempt to rise to the challenge of giving an adequate description of the everyday phenomena of reflective agency and self-control. It is to the prospects of meeting this challenge that I now turn.

## Intentions

Acting in accordance with what I take to be the reasons for acting in a given situation can seem to have little to do with what I desire, but more importantly, can be enough of an act of genuine self-determination to make the Humean account seem like it must seriously overreach. *Prima facie*, both forming an intention and acting in accordance with a previously formed intention seem like they can be examples of this. This is important for present purposes because being addicted seems to involve an impaired ability to do just this relative to a particular domain of action.

I take it that we are intimately familiar with *intentions*, and the role they play in structuring and coordinating our actions over time, from everyday life. And it does seem that in ordinary experience our capacity for forming and executing intentions is not constrained by an antecedently given set of motivations. I can form the intention now to make a stir-fry with the green beans currently in my fridge next Wednesday even though neither the performing of the mental act of forming that commitment is *now* the thing I want most to do nor is it guaranteed to be the case that what I want to eat most on Wednesday will be green beans. Nor, crucially, does it seem that my ability to do these things is conditioned by some set of reasons-unresponsive states that I just find myself in. Indeed, it seems that in forming the intention I exercise a capacity which is reasons-responsive par excellence.

Two features of intentions are especially relevant for our purposes. First, they are intrinsically motivating states. When I act on an intention, I do not need to conjure anew motivational resources for performing that action; the intention motivates me directly. This is what allows my action on an intention to be unconstrained by my desiderative profile (on any ordinary understanding) at the time of action. Second, intentions are closely related to, but importantly distinct from, normative judgements about what I ought to do. We can see this by considering a case where the two come apart: Suppose I am in the course of planning and executing some nefarious activity when my (more virtuous) friend comes along and points out to me the moral pitfalls of following through with my plan. I may accept that her judgement is superior and, in effect agree with her in her making of that judgement, without allowing that to dislodge my plans to carry through with my mischief. Indeed, any example of *akrasia* would seem to make the same point. There is a sense in which the moment of the judgement and the formation of the intention are distinct, but in a normal case, where I succeed in intending to do (and perhaps also, in doing) the thing that I judge ought to be done those moments are pretty much simultaneous. It can seem to be a feature of the virtuous agent that, for her, these moments are only notionally distinct. When all is going well, the function of a future-directed intention is simply to 'preserve for later' that force of the reasons that I judge here and now to bear on that course of action in question.

This is what allows my formation of a future-directed intention to be unconstrained by my desiderative profile (on any ordinary understanding) at the moment I form the intention.[19]

A clarification is in order here: I am not claiming that the capacity for forming intentions is in any frightening sense *radically* free. Of course, it may be that nothing is like that. I am simply claiming that it is implausible to suppose that my capacity for forming and following through with my intentions is constrained by the set of my antecedently given desires (in any ordinary sense of that word) which I happen to find myself with. Of course, I can't simply intend to do *anything at all*. Some limitations will be familiar and non-motivational: I have limited capacities for representing the world, for keeping things in mind, for seeing practical and theoretical connections between things, and so on, which limit the range of things I could possibly get myself to intend simply by limiting the range of things I could even conjure up in thought in connection with one another.[20] However, there will be other limitations which we might choose to characterize as motivational, but which don't very naturally fit the mold of a Humean desire. For example, it is plausible to think that I can't get myself to do (or to intend to do) something which I see *nothing* at all in favour of doing. But to characterize a situation like this as one in which the barrier between me and my action (or my intention) is constituted by a lack of desire, just seems to get the facts wrong. We can and do move ourselves towards a particular course of action because we see something in favour of that action *despite* the fact that we have no desire to do so. If the Humean is simply pointing to fact that there are limitations, decisively beyond my control, on what I can and cannot get myself to do, then I suppose I have no quarrel, but it does seem that these limitations have little to do with the ordinary understanding of desire or motivation, and that to identify these limitations by pointing to something we call the agent's desiderative profile would be to unnecessarily obscure matters.

Indeed, the correct understanding of why someone formed an intention may involve thinking of it in terms which directly contrast with talk of desire and motivation; my forming the intention to use the green beans may have been in the first place a way of tying myself to the mast, knowing that any inclination to do the thing in question that I may be lucky enough to find myself with at the moment (if such there be) almost certainly will have faded when the time comes to act. Intentions have a way of 'preserving for later' the force of appreciating, here and now, that such-and-such is to be done. Moreover, they do so with a particular kind of stability. Richard Holton, echoing Michael Bratman, characterizes this feature thus (Holton 2009, 2-3): 'Stability [as a feature of intentions] is best understood as a shift in the threshold of relevance of information: some information that would have been relevant in forming an intention will not be sufficient to provoke rational reconsideration once an intention has been formed.'

The Humean, it seems, must say that *some*, perhaps unordinary, desire is present at

[19]The fact there is this space, in the non-virtuous agent, between judgement and the formation of the appropriate intention, is part of the reason why I prefer to think of the failure to form that intention as a genuinely volitional failure. I will return to this point again when discussing Wallace's view, below.

[20]Just for example, intending to $\phi$ by $\psi$ -ing is a common phenomenon. But in order to do that, I must have the ability to see some connection between $\phi$-ing and $\psi$-ing.

the time when I execute the action that I earlier formed the intention to perform. It is open, of course, for him to claim that I simply do, my introspective self-understanding notwithstanding, have the desire to do the thing in question. But as I have already noted, there is something at best highly artificial, at worst misleading, about this response. The most plausible way to couch such a desire in the face of the fact that I simply *don't feel like doing that thing now*, would be to appeal to something like a 'desire' to follow through with my intentions, perhaps with some caveats allowing for rational revision. This 'desire' isn't something that I should *expect* to have much of a phenomenological component to it, the Humean may insist, because, after all, it is a very different kind of desire from the desire for a glass of grapefruit juice or a cigarette. It is highly general, something more like a policy than a felt inclination, and more explicitly concerns the regulation of my own conduct than 'ordinary' desires do. But the more detailed the explanation of this difference between my 'desire' and the more ordinary desires becomes, the less it seems like a desire at all. Indeed, the more we characterize it as a policy by which I seek to regulate my conduct, the more it seem like a normative judgement or, indeed, an intention to follow through with my intentions.

But perhaps there is a slightly more sophisticated reply available. Part of what is going on when I follow through on my intentions is that I am exercising my capacity for self-determination. I decided earlier that I should do something in the future, and then, when the time came, I did it. Part of the challenge to the Humean is to make sense of how this is so much as possible if my actions are simply determined by the strengths and vectors of the various forces of which I find myself a patient. The challenge to the Humean thus appears to be a demand for the vindication of these capacities for self-determination, the specifics of the example and the particular competing conception of agency notwithstanding. Understood in this somewhat more general way, the Humean can try to make room for a kind of self-determination even within the meagre resources of his theory. Although our actions are subject to the strength and directions of the various fields of motivational force in which we find ourselves, we also have the power to partially determine which impulse will 'win out'. The strength of a given desire is partly conditioned by elements of the agent's psychology which are, to a certain extent, under her control, e.g., precisely which concrete things would seem to be made better by the satisfaction of that desire (considerations of seeming benefit), or how difficult or easy the desire *seems* now to be to satisfy (considerations of seeming cost). A desire's motivational force may be diminished by focusing on the costs of satisfying it (in terms of the satisfaction of other desires, of course) or by keeping out of mind the benefits of satisfying it. Conversely, the motivational force of a desire may increase the less costly and more beneficial it can be made to seem. By directing his attention (which is under voluntary control) towards or away from considerations which change how things seem to him which may serve to diminish or to amplify the motivational force of the desire, the agent can alter the subjective motivational force that the desire is able to exert on him.

This is a coherent picture. But it is far from clear that these are the only tools of self-regulation that we have at our disposal. Of course, desires in the phenomenological sense can be very relevant for my practical deliberation and I can affect the strength of that felt

motivation by the selective direction of my attention. But it just seems incorrect to suppose that this is all we can ever do to direct ourselves towards or away from objects or courses of action with respect to which we may have highly complex attitudes of valuation. In particular, it seems that whether or not I have formed the intention to perform a certain action, and whether that intention has remained stable (in addition to whether, stable or not, I judged that the reasons in favour of revising my intention are strong enough that doing so would be rational) can be a decisive factor in determining whether I perform that action or not, and it simply doesn't seem that this happens via a process of motivational self-manipulation.

Thus, I think that there is ample room in our moral psychology for states other than beliefs and desires. In particular, forming and following through with intentions seem to be important volitional phenomena.[21] If we are willing to accept this much, I suggest that we think of forming and following through with intentions as capacities of the *will*. I have been suggesting that the Humean view, according to which the will is ever-present but never a cause — or better yet, on which the will doesn't really exist — is too impoverished to capture our ordinary understanding of agency. However, I am sensitive to concerns about metaphysical extravagance, so I have been careful to introduce a distinctive capacity only where I think we need one. There is, for example, no need to shift up to what we might call a 'Cartesian' conception of the will according to which the will is ever-present and *always* a cause — and not just because we aren't dualists searching for the mental cause of every bodily action. The exercise of the will, understood as a set of related capacities for self-control, is simply not required for action. Indeed, most of our actions occur without the effortful intervention of any such capacities. I have suggested that forming intentions to perform actions that desire alone could not determine us to do is one such place where we should locate an exercise of the will, and that following through on such an intention later could be another. There almost certainly are others, but recognizing just these two is enough for us to begin to ask about the ways in which addiction can be a volitional impairment.

We saw above how the Humean position might be understood as the basis for a liberal view of addiction and responsibility, but I think now we can see that it can only do so by greatly misdescribing the phenomenon of agency.

Although I think that both the ability to form intentions and the ability to follow through on them are volitional capacities, I should flag one important difference between them that is relevant for our purposes. Whatever else is involved, forming an intention involves the ability to recognize certain things as reasons. This is the reason why I can't simply intend to do anything at all: there must be something that counts in favour of doing something, from my deliberative perspective, in order for me to be able to intend to do it. Forming an intention has lot in common with, and is in typical cases subsequent to, the making a normative judgement. Following through with an intention, however, isn't a matter of taking there to be reasons in favour of doing something, but a matter of *doing* it.

[21] I am obviously greatly indebted to Richard Holton (2009) and to Jay Wallace (1990, 1999b) for this way of thinking about things.

This is, of course, nothing other than the distinction between recognition and reactivity with which we began this chapter. Understanding the difference between responsiveness and reactivity as a difference (at least) between two different sorts of volitional capacities allows us to see that there are two different ways in which addiction might impair the will. I will consider whether addiction might involve a failure of recognition first, before going on to discuss failures of reactivity.

## 5.4 Failure of Recognition

Let us consider an addict who has not yet appreciated the considerations in favour of abstaining, although there are many. Is there a sense in which she suffers from a volitional impairment, a failure to recognize and appreciate reasons? This is a bit of a tricky issue, since it is normally assumed that the capacity to respond to reasons is a global capacity of rational agents. And as I mentioned in the introduction to this chapter, it would be most implausible to suppose that addiction somehow undermines this capacity *tout court*. Some people who are suffering from addiction can rather seem like they are suffering from a localized failure to recognize precisely the reasons in favour of their quitting. As evidence for the harmfulness of their behaviour to themselves and others mounts, the addict can seem somehow bizarrely impervious. It can take a traumatic event, or some other way of precipitating ('hitting bottom') the badness of the situation to get an addict even to the point where he can even *form* the intention to quit, much less face the difficulties that attend following through. Is this best thought of as a local impairment to reasons-responsiveness which impairs the will? Can we make sense of such a local impairment?

We might be skeptical. But it is hard to know how much of a challenge is presented by the possibility that someone could fail to 'see' or 'appreciate' considerations which are often *literally* all around and right in front of them, without knowing a little bit more about what me mean by 'seeing' or 'appreciating'. If it were literal seeing at issue, we might distinguish between 'seeing' and 'seeing-as' and say that an inability to see something could be traced either to a failure in the visual system akin to a sort of blind spot (failure of seeing) or to a higher-level cognitive impairment of some kind (failure of seeing-as). So perhaps too we can distinguish between an inability to see, in some more literal sense — perhaps to see, or otherwise sense, the facts that reasons consist in — and an ability to see such facts *as* reasons, with their typical normative force, as reasons that I could act on, reasons *to form an intention*.

I think it is pretty clear that any inability to 'see' reasons at work in addiction is not likely to be of the first, more literal kind. Surely it would be extreme to suppose that addicts, even at the limits of depravity, literally lose the ability to see facts. But the second does seem to have some plausibility to it. It is tempting to describe someone who is surrounded by evidence of their condition, who nevertheless does not seem (yet) capable of appreciating that very condition, that somehow that evidence is just not showing up for them *as* evidence. This is a tempting description, and it would nice to make good on it if we can.

## Self-Deception

I think we are familiar with a kind of failure of rationality which fits our bill quite nicely. When an agent is self-deceived, there is a sense in which she is blind to a host of localized, internally related epistemic considerations, without there being any temptation to think she has lost the ability to appreciate epistemic reasons globally.

There are two kinds of epistemic insensitivity that might be work in addiction that might constitute a failure of recognition, both of which might be explained by self-deception. These two kinds of epistemic insensitivity correspond to two different ways in which addictions can be borne by those who have them. Let me explain.

Simply *having* an addiction is one thing. Perhaps having acquired a certain non-natural appetite (see §5.5 below) suffices for this. Many people have addictions (to substances like alcohol, caffeine, nicotine, methadone, or even other opiods) which do not disrupt their lives even though the use of those substances is a rhythmically fixed point within those lives.[22] There is no doubt that there are many happy productive people with addictions to these substances who would suffer, perhaps quite greatly, if they were unable to continue use in a regular way. However, what is typical of the those with this kind of addiction is that their use of the substance is subordinate to other valuable ends. A drink — even a regular one — may conduce to conviviality; an espresso might help one to leap into the day or to extend one's hours of joyful productivity. And so on.

On the other hand, there is what we might call *living the life of an addict*. Typically this happens when drug use becomes an end itself and begins both to crowd out other valuable pursuits — sociality, creativity, etc. — and to subordinate the addict's ordinary organizational activities. We all must engage in the humdrum to be able to pursue what we value. When drug-taking monopolizes what we value, however, our everyday pursuits can become more and more ensnared in the service of that end.[23]

[22]Gene Heyman (Heyman 2009, 52) recounts the story of one such woman, Freida, 81 years old at the time she is interviewed, as follows:

> I first started using drugs after I was divorced. I was smoking opium...all I felt was a good feeling. I kept going every night. I didn't think of the danger. I just smoked every night until I got hooked....When I couldn't get opium, I took heroin...I started to use Dilaudid. I lost my heroin connection on the Lower East Side. I got the Dilaudid from doctors. I got my needles from a druggist in the Bronx. He knew me for years. He must have known I was addicted...I entered the methadone program because I couldn't get Dilaudid. I want to stay on methadone. At my age, if I got off I'd die, I'd never make it. I'm happy. As long as I've got money, I can play the numbers. I play numbers every day.

Heyman goes on to say that Freida never voiced any regrets about her drugs use and emphasizes how she reports satisfaction with her life. Freida and other participants in harm reduction programs like methadone programs certainly qualify as having addictive appetites, but it may be far from clear the extent to which they or others are harmed by those appetites.

[23]The use of some psychedelic drugs is interesting in this connection. Many people who have used psychedelic drugs such as LSD and psilocybin mushrooms (myself included) report thinking of those experiences as valuable (even highly valuable) insofar as they conduce to emotional insight and other forms of

The distinction between those who live lives of addicts and those who do not thus marks the distinction between two very different kinds of problems that an addict might have. Addicts who merely *have* an addiction may (or for that matter, may not) face difficulties associated with their habit. For example the habit may come with unwanted financial costs, or costs in terms of one's health. These costs may provide reasons to abstain or to cut back, even reasons that are strong or decisive. Insofar as these factors make continued use fraught, these addicts may be said to have a problem.[24] However their problem is of a different kind (and usually of a different degree of severity) than one who is living the life of an addict. To live the life of an addict is to make a *valuational mistake*. It is not only to assign to a pursuit a status it does not warrant[25] it is to allow that pursuit to crowd out other more valuable pursuits, leaving one with a life which is comparatively impoverished.

The distinction between these two different conditions that addicts find themselves in is thus drawn evaluatively. It is also vague. But both of these things are as they should be. The extent to which someone can be said to be living the life of an addict is pretty clearly a gradable matter: there is variation in the degree to which the addictive pursuit has in fact monopolized the addict's energy and attention and there will be some amount of unclarity concerning whether some cases exhibit enough of the relevant kind of misvaluation to count. But drawing this boundary precisely is not what is of importance here, nor should it be expected. What matters is that one's concerns and attention can be more or less greatly dominated by the addictive pursuit, and that when this happens the addict is committing a valuation error of corresponding severity.

Addicts in either situation can fail to recognize the situation that they are in. Of course, this failure is only potentially an excuse when it is not itself something for which the subject is responsible. One way such a responsible failure often manifests is as self-deception. Some explanation has to be given for why such strong and clearly available evidence goes ignored by

---

self-knowledge. There is, of course, also a way in which, *as* a valuable experience, other activities can be made valuationally subordinate to it, but given how severely debilitating and overwhelmingly stimulating such experience can be, there is really no possibility of a recognizable life *devoted* to, and filled beyond saturation with, episodes of such drug use. It is possible that this partially explains the clinically very-well attested nigh-impossibility of becoming addicted to such drugs. (Other factors such as rapidly and exponentially increased tolerance — such as to make repeated use almost completely ineffective — surely also play a role.)

[24]Nicotine users perhaps best illustrate the potential dangers of even some addictive behaviours that fall short of living the life of an addict. There is no doubt that smoking cigarettes comes with a great many health hazards and nicotine is one of the most addictive substances habitually used by human beings. But it is very rare to hear of someone whose life was *devoted* to cigarettes. This is perhaps in large part because of the comparatively low cost and high availability of tobacco products and the comparatively low social stigma associated with nicotine use. When drugs are expensive or otherwise difficult to obtain and when addicts feel driven to use in secret, the increased overall effort associated with maintaining one's habit increases greatly the likelihood that other valuable pursuits will have to fall by the wayside in order to sustain the habit. This is perhaps one of the most powerful arguments (among the many that I know of and endorse) against policies of drug prohibition.

[25]For an elaboration of the idea that what is *wrong* with addiction is that it is devotion to an unworthy object see Summers 2011.

addicts, and self-deception provides this explanation. This does not rule out that there could be failures of recognition that are not self-deceptive, and which might constitute excuses. But here I wish to focus on self-deception.

Addicts who merely *have* a problematic addiction need only deceive themselves about the seriousness of the costs associated with continuing their habit. This kind of irrationality is indeed almost paradigmatic of self-deceptive thinking. For example a moderate but longtime cigarette smoker may think, of the relevant health risks: 'That will never happen to me'; 'I can handle it'; 'The science on that isn't settled'. The evidence he must ignore comes not only from scientists, health agencies, and his health-conscious friends, but from his own experience as well. Ever-decreasing energy levels, frequent pulmonary infection, and elevated blood pressure will also confront him. And again he may believe: 'That is just getting older for ya'; 'It's been a bad winter'; 'It runs in the family anyway'. However, if the overarching belief, viz., 'I am fine', is something to which the addict is sufficiently strongly committed — for fear of facing up to what it might mean if it were false; for an unwillingness to make the necessary changes; or for any other reason from which he derives his present preference to continue believing it — he will try to ignore or re-interpret all of this evidence. Doing so does not require any paradoxical doxastic or motivational states. Nor is there anything particularly puzzling about attributing to him the overarching belief that his pursuit is not — in general or in any particularly threatening way — problematic. I do not know whether it is plausible to think that such a belief could come about from the operation of a DARM. It seems to be an empirical question whether there is any mechanism keyed to such an outcome. But even without it, Self-Deception as Omission* captures what is going on here. Our smoker believes that he is not in trouble, however that belief came about precisely, and he omits to seek evidence against this belief and to appreciate that which confronts him, both because he desires that it be true that he is not in trouble.

For addicts that are living lives of addicts, things are a little more complicated. Here it will be helpful to invoke Finagrette's account, and to again highlight some of the affinities that my account has with his. In order to continue believing she is ok, an addict of this kind has to ignore or fail to appreciate the very valuational structure of her life. This is evidently a more elaborate form of self-deception, and it befits Fingarette's language nicely (Fingarette 1969, 62):

> [T]he self-deceiver is one whose life-situation is such that...he finds there is overriding reason for adopting a policy of not spelling-out some engagement of his in the world...The consequence of this is that he may be observed as one who is in fact engaged in the world in a certain way...; yet he is unable (by virtue of his commitment) to spell this fact out to himself or to anyone else. Thus when the issue is raised, he does not, cannot, express the matter explicitly at all. He is in this respect in no better position than anyone else. He tells us nothing but what he tells himself.

On my view, the reason for 'adopting the policy of not spelling-out' is a preference for

the truth of some proposition, one which, as the case may be, is at the center of one's self-understanding. Fingarette, however, prefers to eschew what he calls 'cognition-perception' vocabulary (the language of belief, knowledge, evidence, etc.) (Fingarette 1969, 66–67):

> The self-deceiver is one who is in some way engaged with the world but who disavows the engagement, who will not acknowledge it even to himself as his. That is, self-deception turns upon personal identity rather than the beliefs one has.

But I don't think what one believes and one's personal identity are separable in the way that Fingarette is imagining. Now, obviously Fingarette is not talking about one's *numerical* identity here, but rather something more akin to one's narrative self-understanding. But one's narrative self-understanding *is* a matter of what one believes, viz., what one believes about oneself and about one's relation to the world in terms of what one values. I have a conception of myself as belonging to certain social and cultural groups and to a certain time and place; as standing in certain valuable relationships and engaged in certain valuable activities; as someone with a certain character and with certain commitments, and so on. These things are *beliefs* that I have. I have come to hold them in response to a lifetime of experience as the person that I am; they are responsive to evidence and change over time; and they exhibit a degree of coherence with the other things that I believe. For example, my conception of myself as a philosopher and a member of a certain professional community depends not only for its content but also for its rationality on my other beliefs about what kind of activity philosophy is, who my friends and acquaintances are, how I spend my days, and so on. I cannot believe that I am a philosopher unless I believe all of these other things. But the dependence also goes the other way. Since *what it is* to be a philosopher just is to live a life like the one I am actually living, to believe all of the things that I do believe about my own life exerts a great deal of rational pressure to believe that I am leading the life of a philosopher. This mutual dependence between the identificatory state and my other belief states is best captured by thinking of the identificatory state as itself a belief state.

Fingarette is right, however, to give one's personal identity a central place in his discussion of self-deception. Self-identificatory beliefs are excellent candidates for beliefs that one may hold self-deceptively. This is so for at least two related reasons. The beliefs that we hold about ourselves and about the fundamental aspects of our identities are near to the center of our webs of belief: a lot depends on their stability. This makes them inherently resistant to revision, just on the grounds that to revise would set off a cascade of costly revisions throughout the belief network. But human beings also have a well-known bias towards positive self-assessment.[26] This makes it all the more likely that what we believe about ourselves is positively valenced — or any rate, more positively valenced than the evidence, strictly speaking, warrants.

[26]There is a large family of egocentric biases, but the most well-documented is the self-serving bias: subjects are more likely to engage in internal causal attribution for successes, and external causal attribution for failures.

In §4.2 I talked about being 'emotionally entangled' with a belief as a way of cashing out what it might mean for someone's desire that it be true to play a role in his continuing to hold it. I said that for someone to desire that $p$ is for the truth of $p$ to bring the world closer into conformity with how, from his perspective, the world ought to be. Self-identificatory belief will almost always be like that. It will also, obviously, be a belief that the subject already holds. Thus, the conditions are ripe for Self-Deception as Omisson* to be satisfied.[27] We must simply find a willful failure to seek, recognize, or appreciate evidence.

In addition, such a failure seems precisely what we have need of in order to capture a certain aspect of the behavioural syndrome involved with serious addiction. One of the most striking features of addicts in this situation is how impervious they are to reasons whose force is obvious to everyone else. Therefore, there *is* a failure to seek, recognize, or appreciate evidence. The only question remaining is whether it is willful. Here it is obviously not possible to say that it always is willful. But let us consider the self-identificatory belief that is the likely culprit here: the belief 'I am not an addict', or indeed: 'I am not living the life of addict' (or something that entails this). The motivational reasons that a subject likely has for hanging onto that belief are obviously quite strong.

As I argued in §4.2 and §4.4, when someone is self-deceived they are attributability-responsible for their self-deception and, typically, also blameworthy. The very idea of self-deception as something willful seems to require that it be attributable, but whether the self-deceiver is blameworthy depends on whether, like B.X., she has a weak excuse. There is no perfectly general answer to this question when it comes to addicts. It is of course not impossible that someone in this situation could have good reason for shirking a standing epistemic requirement. But the more flagrant and more serious the violation of the epistemic norms — or to put it somewhat differently: the more epistemically *vicious* the conduct — the harder it will be to find an excuse that fits the bill. Thus, those who are living lives of addicts and are self-deceived about it are much less likely to find themselves excused than those who are merely self-deceived about the seriousness of their moderately vicious habit. For example, tobacco use is significantly above average in inpatient psychiatric populations. However, a lot of these patients find tobacco use therapeutic. (Often caretakers will even encourage continued smoking if it provides familiarity while the patient goes through other uncomfortable changes associated with treatment.) If we assume that they have the same information about the health risks associated with smoking as everybody else we might think that there is some sort of willful failure to appreciate that evidence. And such there may be. But, if smoking is helping with the treatment of a more serious (or at any rate, more immediately serious) condition, any self-deception about the risks of the smoking might be justified, especially if the condition they are suffering from, and the recommended course of therapy, are acute and arduous in such a way that greatly taxes attention and other epistemic

[27] Whether Self-Deception as Omission, the view from Chapter 3, will also be satisfied will depend on whether we should class the suite of cognitive processes that are responsible for producing the self-identificatory beliefs as type-1 processes. At present, I don't know how to answer this because it is at best unclear how such beliefs are *formed*. As I have been emphasizing, they are are developed slowly over time and take up pride of place in the agent's web of belief.

resources.

However, the more serious the addictive condition becomes, the more pressing becomes the need to acknowledge the evidence, the more difficult it becomes to ignore or reinterpret it, and the more willful the failure of recognition must become.

Thus, where a failure of recognition in addiction is due to self-deception, it may or may not be something for which the addict is blameworthy. Correspondingly, it may or may not provide an excuse when assessing whether the addict is blameworthy for continuing the addictive pursuit. Nevertheless, it is plausible that blameworthy (non-excused) self-deception is involved in the more serious cases of recognition failure in addiction. If this is correct, there is an important sense in which addicts collude to reduce the resistance to the continuation of their addictive pursuit. It is perhaps for this reason that it really is an important first step on the road to recover to announce somberly to a room full of strangers 'I have a problem'.

## 5.5 Failure of Reactivity: Willpower

I want now to consider the sense in which addiction may involve a volitional failure which is a failure of reactivity. I will begin with R. Jay Wallace's discussion along these lines as a jumping-off point. Wallace shares the aim of rejecting the Humean view of action in order to make room for the exercise of distinctive volitional capacities. He complains, I think rightly, that the Humean picture (Wallace 1999a. 633)

> leaves no room for genuine deliberative agency. Action is traced to the operation of forces within us, with respect to which we are ultimately passive, and in a picture of this kind genuine agency seems to drop out of view.

Wallace recommends what he calls 'the volitional model' over the Humean model, on which we should (Wallace 1999a, 636)

> acknowledge a third moment irreducible to either deliberative judgement or merely given desire. This is the moment that I shall call volition. By "volition" here I mean a kind of motivating state that, by contrast with the given desires that figure in the [Humean] conception, are directly under the control of the agent. Familiar examples of violitional states in this sense are intentions, choices, and decisions.

I have been emphasizing the role of intentions, but I think Wallace and I are largely in agreement about the need to include distinctive volitional capacities in our picture of agency, and broadly speaking, the features which I think they have, including, when all is functioning well, being directly under the control of the agent. With this in hand, Wallace then proceeds to identify a particular feature of the addict's predicament which interferes with his ability to effectively exercise these capacities: the character of the desires to engage in addictive behaviour (which Wallace calls 'A-impulses') that the addict is subject to. I also

think Wallace is right to emphasize the role played by these desires; any picture of addiction that failed to recognize the role of desires in the persistence of addiction would surely be incomplete.

On Wallace's way of thinking about desires, they are essentially phenomenological, and quasi-perceptual, representing a concrete course of action as in some way (perhaps highly) pleasant. On this way of thinking about them, desires do not determine our actions with their relative strengths, as on the Humean view, but can nevertheless present real obstacles to acting well. This is primarily due to their unresponsiveness to our reflective judgements, or what Wallace calls their 'unruliness'. It was one of my primary complaints against the Humean view that it problematically slides between the phenomenological and the pro-attitude sense of 'desire' on the way to the conclusion that our actions ultimately issue from states that are unresponsive to reason. However, I do agree that desires in the phenomenological sense (at least the desires at work in addiction) *are* largely unresponsive to reason, and this allows us to see how they might, if strong enough and persistent enough, constitute a volitional impairment.

Wallace's idea is: Given the peculiar strength and persistence of A-impulses, relentlessly representing the drug-taking behaviour as highly pleasant and highly salient, they can interfere with the agent's ability to do what reflective judgement reveals would be best even if we assume that they are not interfering with her ability to arrive at such a judegment.[28] He puts it this way (Wallace 1999a, 648):

> The A-impulse that persists in a situation of this kind is the extreme case of the phenomenon of temptation, a psychological condition that facilitates the choice of

[28]Wallace is willing, of course, to concede that A-impulses might also interfere with the agent's ability to arrive at that judgement by impairing, for example, her ability to keep firmly in mind the considerations which bear on that judgement. But, interestingly, the way he is thinking of the will, this would not constitute a volitional impairment, but, instead, an impairment of rationality alone. He says 'A defect of the will...would need to be a form of interference with the processes of reflective agency which go beyond, and are independent from, deficiencies in respect to rationality alone' (Wallace 1999a, 636). In our terms, it seems that Wallace does not consider a failure of recognition to be a volitional failure. This does seem sensible: if recognition concerns merely an agent's ability to recognize the force of reasons, there seems to me nothing wrong in saying that this recognition is the function of judgement and that, therefore, any impairment in the ability to do so should be located exclusively within the purview of the norms of rationality. But the ability to form intentions requires the ability to appreciate reasons and make judgements. And since I have been conceiving of the ability to form intentions as a volitional capacity, it seems that there might be some room for thinking of an impairment to recognition as *ipso facto* a volitional impairment. Therefore, I prefer not to rule out this possibility. Wallace himself concedes that the distinction between rational and volitional impairments might not itself be particularly important: 'In the end, admitting that there are norms of rationality governing volition as well as belief deprives the distinction between defects of rationality and defects of will of its theoretical interest. The relevant distinction to draw in this area is within the class of defects of rationality, between impairments of our capacity for practical judgement and impairments to our capacity to choose in accordance with our practical judgements' (Wallace 1999a, 651). The difference I wish to mark out is simply to try to make room for the idea that a certain kind of impairment — impairment of recognition — could underlie both an impairment of judgement and of the ability to form the appropriate intention *because* the latter depends on the former.

> an action that the agent believes is ill-advised, by directing the agent's thoughts onto the alleged attractions to be gained through that action.

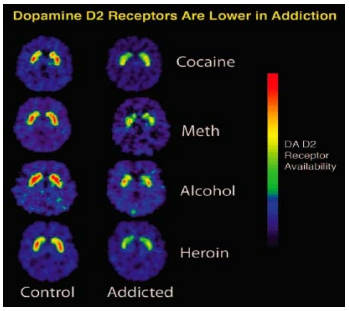
This sounds quite plausible. Furthermore, as addictive behaviour becomes more and more entrenched, both at the level of behavior and at the level of the brain, it also becomes more and more *automatic*. And the more automatic it becomes, the more the deployment of the will will be required to overcome its inertial force. Consider how much willpower must be required to successfully sustain a hunger strike. Seeking food when hungry is such an automatic response that most us can, almost literally, do it without thinking. Perhaps this is as it should with the 'natural' appetites. But we also have the capacity to acquire an almost unbelievable range of activities as habits and to elevate them to a very high level of automaticity. At the extreme, they can become things that we literally do without thinking. And complexity seems to be little or no barrier: for a practiced chess master the extremely difficult task of finding a strong (or perhaps the optimal) move in chess can approach this level of automaticity. When the task is not harmful and requires considerable skill, we are likely to exhibit a certain appreciation for it. But if the task is banal, harmful, or simply unworthy, we may express disdain or other forms of disapproval. The complexity of behaviour involved in highly cemented addictions can be astonishingly large, but doesn't exactly exhibit what we would call *mastery* — perhaps partially because that term has an uncompromisingly positive valence, or perhaps because none of the individual activities constituting a pattern of addictive behaviour need to exhibit much *skill*. For all that, that doesn't prevent such behaviour from becoming highly automatic over time. As an Aristotelean might put it, we are blessed with having a second nature, but that blessing also imposes a burden: we become practiced in what we practice and it therefore behooves us to be careful what we practice. With enough practice engaging in the addictive pursuit can become that which the agent will do, unless she or something intervenes to prevent it: it can become a default action.[29]

There is thus a considerable amount of resonance to Gary Watson's idea that addiction is an 'acquired appetite'. Appetites aren't just desires. Appetites structure patterns of behaviour by producing various degrees of discomfort when the subject is deprived of the object of the appetite for a prolonged period of time. Appetites don't simply 'go away' on their own, and they produce pleasure when they are satisfied. As Watson puts it (Watson 2004a, 76):

> Appetites involve positive and negative inclinations. We are naturally hooked on food and drink. When I am hungry, I typically become more or less uncomfortable. That is distracting. I desire to various degrees to relieve this discomfort, but that is not all. More positively, the distinction between the edible and the nonedible in my environment becomes highly salient to me. Depending on experience, certain sorts of food are especially alluring and their consumption intensely enjoyable.

[29] In order to fully make good in this claim, given the way I defined default action in §2.4, I will need to appeal to the concept of willpower as a System 2 capacity. For that, see below, §5.5.

What makes addictive appetites different from the natural appetites for food and drink and the like is that they are acquired over time by a history of behaviour which 'forges a cognitive and motivational link' between the substance and the pleasure (and increasingly, relief from distress) it produces. There is thus *something* to the dominant picture of the *etiology* of addiction. Most of us will be familiar with images like this:




The idea is supposed to be: Repeated overstimulation of dopamine receptors leads to lower baseline dopamine receptivity in people with a long history of substance abuse. These are plausibly the neurological correlates of acquiring the addictive appetite: the same behaviour that forges the link between the drug-taking and pleasure also leaves tracks at the level of the brain. They are deeper and perhaps more 'damaging' than the 'tracks' left by, say, ten thousand hours of tennis practice, because they involve *over*stimulation and result in lower baseline receptivity. But this should be largely unsurprising. Not only is it true that we should expect any activity which is practiced enough to show some such neural correlates, the particular nature of what is being practiced — the acquiring of an appetite — should leave us to expect to find it showing up in the brain's reward system, and it is quite plausible that these observed changes underlie and explain the peculiar strength and persistence of A-impulses, as signs of such an acquired appetite. Once an addict has acquired an appetite in this manner, they face a very difficult situation once they resolve (form a strong future-directed intention) to abstain. They now must mobilize their volitional resources to oppose a bodily inclination which has acquired the urgency of a natural, life-sustaining appetite, the satisfaction of which has acquired the ease and quasi-automaticity of a well-practiced skill.

But there is something that I think Wallace's view leaves out. There seems to be a capacity, of varying strength across individuals, to effectively stick to one's intentions, and being weak in this respect might put someone at higher risk for sustained addictive behaviour, even after he has realized that he should quit. It is of course true that, all things being equal, the strength of interfering factors, such as temptations, will negatively impact an agent's capacity to do so. But there seems to a second kind of factor, one more directly contributed by the agent by way of something that he *does*, which is independent of the strength and character of the impulses he might resolve to resist. This capacity corresponds fairly well to the ordinary concept of willpower. It is the concept of an executive capacity, the capacity to stick to one's intentions. Some people seem to be better at this than others. As Richard Holton puts it, '...having formed the resolution not to be moved by certain desires, they

are better at acting in accordance with it, and at turning the corresponding intentions into action' (Holton 2009, 130).

I think that allowing the concept of willpower into our moral psychology has a lot of advantages, and it fits naturally with the conception of intentions I have been working with. Indeed, I have throughout been talking of *both* the forming and the following through on intentions as volitional capacities, and willpower is nothing other than that second ability. It is most certainly a volitional capacity, but it deserves special recognition. If we simply added the ability to form intentions to an otherwise Humean psychology, we wouldn't have gotten ourselves far enough away from the shortcomings of the Humean picture. Supposing that there are motivationally efficacious states whose formation is directly under the control of the agent won't do us much good if they are simply thwarted whenever there is a countervailing desire of sufficient strength. Worse still, the account seems to get the phenomenology wrong. The Humean picture augmented with intentions must say that when we successfully act on our intentions it is because the motivational strength of our intentions just happens to be stronger than that of any countervailing desire we happen to find ourselves with. But this is surely not what it *feels* like. It feels like a struggle to stick to your intentions in the face of countervailing desires. In the case where you succeed, it doesn't seem that this is just because you have managed, somehow, at the outset, to imbue a state with enough impetus to coast relatively undisturbed through a field of forces pointing away from where you intended to send it.

Rather, what seems required is a capacity for *resisting*. It has been no part of my critique of the Humean picture that desires *don't*, as a matter of fact, have the power to move us all on their own. Indeed, one can imagine a particularly unreflective existence which consisted in not much more than being led about in this way. This suggests that the exercise of willpower is the exercise of a capacity to intervene and attempt to prevent from happening (with varying degrees of effectiveness) what would happen otherwise due to automation, viz., to avoid a life dominated entirely by default action.

Thankfully, this capacity has been fairly effectively operationalized in experimental psychology. Recall once again the distinction between System 1 and System 2. At first this division was devised to account for certain experimentally observed biases, but the framework has proved to be much more useful than that. There is a kind of easily observable phenomenological interaction between the two systems. System 1 whirs and grinds away, typically offering up plausible intuitive verdicts to problems such as the bat-and-ball problem. But occasionally it gets things wrong. You aren't stuck with the error, you can overthrow it, but there's a catch: it takes *effort*. This is one way in which conscious effortful activity opposes itself to what would happen otherwise. Without this exercise of System 2, you find yourself with a false belief.

The bat-and-ball example is trivial in that the stakes are very low, the 'strength' of the 'forces' you need to oppose if you're going to get it right are comparatively weak, and doing your arithmetical duty is on the whole relatively easy. But it seems that the capacity that we have for opposing the various subpersonal, and indeed, external forces that act on us, produce attitudes in us, and seduce us, is both singular, and limited. As Kahneman puts it:

'Self-control and deliberate thought apparently draw on the same limited budget of effort' (Kahneman 2011, 40). Indeed there is mounting experimental evidence that one's ability to carry out these effortful executive tasks is not only limited, but that it is something of a skill. That is, how much self-control someone can exert, or how good they are at sustaining effortful System 2-type thought, is an ability that varies across persons and can be trained to higher levels over time within a single person. These are all features of what I shall call willpower. Let us consider some of the evidence that this capacity varies across individuals before considering some of the evidence that all effortful activity draws on the same store of mental resources.

Evidence that the capacity for self-control (the ability to stick to one's intentions, in our terms) varies across persons comes first from a developmental perspective. In his illuminating summary of work on the topic, Walter Mischel (Mischel 1996) notes that the ability to delay gratification for a larger larger reward develops in children around the age of four or five. By the time they reach age six, almost all children have it, but to varying degrees.[30] In experiments such as these, children are told that they could have something tasty, say, a cookie or a marshmallow, at any time, either because it was left out for them, or by ringing a bell to have someone come and deliver it, or that they could have a larger reward — say, two cookies or two marshmallows — if they could wait until an adult returns to the room. As one might expect, differences in performance were found between the groups of children who were left exposed to the tasty treats compared to those who were not. Something which presents itself, literally in front of you, can be a much more difficult motivational obstacle to overcome. But the most striking results for the hypothesis that the capacity for self-control is of variable interpersonal strength comes from the results the experimenters obtained from within the group of children who were left exposed to the rewards. As Mischel puts it: 'This condition created a situation in which the individual differences in the ability to cope with this frustration [generated by the exposure condition] should be activated and visible' (Mischel 1996, 210–211). One of the aims of this work was to try to establish a correlation between self-regulatory ability at a young age and positive social outcomes later in life. Seen in this light, Mischel expresses the rationale for focusing on the group in the exposure condition rather than the obscured condition: '[D]elay in the reward-obscured conditions was not expected to to be diagnostic of self-regulatory ability, since delay in that condition was not particularly difficult or frustrative for any of the young children we studied' (Mischel 1996, 211).

When the results from this group were compared in longitudinal studies with the participants' verbal and quantitative SAT scores, a high positive correlation was found between self-regulatory ability as measured in experimental conditions and high test scores. The researchers conclude that there is something akin to a personal trait that they had measured early in life, which predicted a wide variety of social and academic outcomes.[31] But

[30] Although not fully developed at this stage, further research has suggested that once this ability does become fully developed, it remains relatively stable over a lifetime (Casey et al. 2011 and Mischel et al. 2011).

[31] Strong self-regulatory ability also correlates with, e.g., staying out of prison.

a skeptic might ask: 'Have they really found that?' In particular, we need it to be that the differences found are differences in the ability to control oneself, keeping fixed the degree of distraction or temptation presented by the tasty treat, if we are really to have a found an independent sort of volition. For example, couldn't it be that some children just experience the temptation more acutely? Wouldn't that explain the results just as well?

I think this is dubious, for a number of reasons. First, this suggestion at first seems plausible because it is masquerading as something that it isn't. Of course it would be experimentally undetectable (I suppose) if some children's desires were simply *globally* all stronger than other children's desires. Insofar as we can make sense of this possibility, it does seem to be something that might avoid empirical detection. But that is not the possibility we are envisioning, since when the children yield to temptation they are pitting the strength of the desire to give in now, against the desire for the larger reward later *plus* the strength of their ability to stick by their intention to wait. But it is the strength of their ability to stick by their intention which is being measured, and the contribution it is required to make in order to allow the child to stick to his resolve is the same even if we uniformly inflate the strength of all the child's other desires. In order for it to be plausible to account for the experimental results in terms of variable strength of temptation rather than variable strength of willpower, it would have to be that the children who performed poorly on the task experience the temptation to give in as stronger (relative to other subjects) without experiencing the desire for the extra reward (factored out from the contribution of the intention, and taken relative to other subjects) as concomitantly stronger. Now, this could be. It could be that, just as temptations that are literally closer to hand exert more motivational force on us, temptations that are temporally more proximate also do so. It could be that these children are representing something good *now* as with more phenomenological force than something which is objectively greater, but also *later*. But if the children who performed poorly were simply representing the *now*-reward as phenomenologically more tempting and *that*, rather than weakness of an autonomous self-regulatory capacity was what explains the experimental results, we should expect that representing the temporally more proximate reward in this way would interfere with the children's ability to *form* the intention to wait — that is, we should expect to observe a failure of recognition, an interference with the ability to appreciate what would really be better about waiting which would lead us to expect to observe no internal struggle. Why would the subjects form the intention to wait in order to get what they take to be a smaller reward?

But that is clearly not what is going on in most of these cases. The children visibly (and very cutely) struggle to resist and this suggests that they have indeed formed the intention to do so, and that their failure is not more fundamentally a failure of recognition.

Further, it is simply not implausible to suppose that the level of temptation experienced by the children is roughly equal across subjects. The objects of temptation (marshmallows or cookies) were chosen to have near-universal appeal to children of the appropriate age, but are nevertheless comparatively trivial. So it would be a little odd to suppose that the results could be explained by some children experiencing a phenomenologically much more forceful pull towards the object.

Related to the thesis that willpower is of variable strength across individuals, is the idea that willpower comes in limited supply, and further, is drawn upon by a wide variety of effortful tasks. Roughly, we can think of the class of tasks that draws upon these reserves at the class of System-2-activities. Experiments designed to show these results typically have subjects perform one kind of effortful (System 2) task, and then check for reduced performance on a different kind of (System 2) task. The sorts of tasks seen in the first group, which Kahneman characterizes as 'involv[ing] conflict and the need to suppress a natural tendency' (Kahneman 2011, 42), are as diverse as:

- avoiding the thought of white bears
- inhibiting the emotional response to a stirring film
- making a series of choices that involve conflict
- trying to impress others
- responding kindly to a partner's bad behaviour
- interacting with a person of a different race (for prejudiced individuals)

And some of the corresponding failures are similarly diverse:

- deviating from one's diet
- overspending on impulsive purchases
- reacting aggressively to provocation
- persisting less time in a handgrip task
- performing poorly in cognitive tasks and logical decision making

'Suppressing a natural tendency' sounds an awful lot like trying to resist an impulse, or trying to prevent what would happen otherwise due to automation; the tasks had to be designed so that subjects could not simply lazily allow System 1 to try to perform them, but instead required the effortful operation of System 2. And we can see pretty clearly what Kahneman has in mind here. Under normal circumstances, subjects would prefer (because it is easier) to exhibit the appropriate response to stirring scenes and images; think of white bears (?); refrain from fawning over, or performing spuriously for, others; avoid tolerating with gentle good humour someone's bad behaviour or spending time in the company of those one irrationally dislikes *because* these things take effort. In the absence of a reason to do these things, subjects won't, and that reason is not likely to be provided by the desire for merriment and the wildly mistaken belief that therein merriment can be found. That is not, of course, to say that the will is required for anyone to get themselves to do anything which

people under normal circumstances might find unpleasant. But if we assume that the subject has no particular desire to do these things, the will may very well be required. But what the evidence shows is that we have a limited and varied ability to make ourselves do things which require such effort. Indeed, the appropriateness of this characterization of the things we have a limited capacity to do, as general as it is, can be seen if we appreciate the following consequence of the hypothesis in question: if what explains diminished performance on the second task is reduced availability of a cognitive resource *because* that very same resource has already been tapped by the performance of the first task, it seems that we should expect results of the same robustness if we swap the manipulated and the responding variables. Given this, it is hard to see what would characterize the group of tasks taken together other than to say that they require cognitive effort to perform. The term 'ego-depletion' has been coined to refer to this weakened volitional state in which the subjects enter the second half of the experiment.

Since the exercise of willpower is to be located in the suite of System 2 processes, the failure to exercise willpower may enable the performance of a default action. Recall that in §2.4 I defined a default action, ignoring external causes, as an action which an exercise of System 2 is required to prevent. What these experiments reveal is just how many different kinds of actions require depletable System 2 resources to perform or refrain from performing, just how many actions are default actions. In someone who has a deeply entrenched addictive appetite, the performance of the addictive behaviour is also a default action.

These results would bolster the position of someone who wanted just to focus on the strength and persistence of A-impulses in locating a volitional impairment implicated in addiction: If every attempted exercise of self-control depletes ones's limited ability to do more of the same, an A-impulse characterized by a very high degree of strength and persistence is likely to present a formidable challenge not just because of its unruliness, but because of its ability to outlast: it is not diminished at every new attempt to resist, but the one who resists is. It isn't just stable, it is as though it fights back. But do these results also provide support for the idea that the strength of one's willpower is an independent variable when considering the psychological situation of someone who has formed the intention to quit?

I think it is very natural to think that these results do support the second claim as well. For one thing, the ordinary conception of willpower seems to be as of an autonomous faculty: we can imagine two subjects facing the same strength of desire to capitulate on an intention, but one person succeeding, and the other failing because, as we might say, the one had more willpower than the other. Insofar as the concept that has been operationalized by these psychologists is closely related to the ordinary concept of willpower (which some, at least, are explicit about trying to maintain), we might expect it to share this feature. But if we are to come to a satisfactory understanding of such an autonomous capacity, we shall have to get clearer about what it is for it be 'stronger' or 'weaker'. It is clear that we want to say that willpower is stronger in someone who less often fails to follow through with what he intends, but of course, there are many factors that might contribute to such a failure, and we need to try to isolate the strength of just one. In particular, there is one important potential confound it is worth pausing to discuss.

Roy Baumeister is fond of drawing an analogy between willpower and muscles (Baumeister 2012). Just like muscles, our wills fatigue over time, and just like our muscles, our wills fatigue in response to wide variety of otherwise seemingly unrelated tasks. The muscle analogy is helpful, but it does have at least one potentially serious drawback. There is some sense in which, at the limits of exertion, muscles literally *collapse*. However, most cases where ego-depleted people fail on subsequent tasks earlier than controls are not cases where they literally could not have persisted a little further, resisted a little longer, etc. Indeed, the capacity for self-control seems to be highly sensitive to rewards (Fingarette 1988, 38).

If all that is required to improve subjects' performance on the second of two ego-depleting tasks is to increase the reward for doing so, it might seem like we have undermined the very thought that what explains the reduced performance in the original experiments was the comparative unavailability of a volitional resource. What is limited about a resource that we can seemingly somehow replenish all on our own if we are given a large enough bribe to do so? It seems that the difference between people who do and do not perform well on the second ego-depleting task is not the difference between people who can and can't, but rather a difference between people who are willing to undergo the required exertion and those who are not. But what would explain this? Might it be that subjects' ability to bolster themselves when offered a reward is explained by their having a *desire* for the reward which is larger than their desire to not undergo the required exertion?

Not only has it been found that addicts can modulate their drug-taking behaviour if given enough incentive to do so, epidemiological evidence (such as that I cited above in §5.2) suggests that this is the natural course which addictions follow: as drug-users get older, they find meaning in alternative pursuits such as a career or a family, and they recover on their own. I think once again the difference between those who mature out and those who do not is not a difference between those who can and those who can't but a difference between those who are willing and those who are not. And again, we ask, what makes this difference?

I think it would be a mistake to think that the difference was simply one of desire. It is not simply that the addicts who mature out of their addictions all have some unreflective desire for a successful family life, or a good career, or whatever it may be, but rather that the value of these other things comes to dawn on them over time, they come to appreciate the reasons in favour of pursuing these things instead of substances of abuse likely to interfere with those things through a process of reflective evaluation which takes place over the course of years. All this is to suggest that the distinction between the volitional obstacles faced by someone who has formed the intention to quit and those faced by the person who has not yet formed that intention is not as clear as I have been making it out to be up until this point. Even someone who has admitted she has a problem and resolved to quit may be unable to appreciate the force of reasons that someone without her addiction might be in a better position to see. Nevertheless, as before, this may be a genuinely volitional impairment, because, as before, it may be due to self-deception. The difference between someone who appreciates the reasons in favour of family life over drug abuse is still a difference between someone who is willing and who is not willing, if the reason for the one's failure to appreciate reasons is itself a volitional failure.

These difficulties aside, there is another feature of the muscle analogy that is quite striking. It seems that subjects can increase their willpower over time with training. In one of the first demonstrations of this idea, Mark Muraven and his colleagues asked volunteers to follow a two-week regimen to track their food intake, improve their moods or improve their posture (Muraven, Baumeister, and Tice 1999). Compared to a control group, the participants who had exerted self-control by performing the assigned exercises were less vulnerable to willpower depletion in follow-up lab tests. In another study, he found that smokers who practiced self-control for two weeks by avoiding sweets or regularly squeezing a handgrip were more successful at quitting smoking than control subjects who performed two weeks of regular tasks that required no self-control, such as writing in a diary (Muraven 2010).

Other researchers (Oaten and Chang 2006) assigned volunteers to a two-month program of physical exercise — a routine that required willpower. At the end of two months, participants who had stuck with the program did better on a lab measure of self-control than participants who were not assigned to the exercise regimen. The subjects also reported smoking less and drinking less alcohol, eating healthier food, monitoring their spending more carefully and improving their study habits. Regularly exercising their willpower with physical exercise, it seemed, led to stronger willpower in all of these other areas of their lives.

That willpower can be increased over time provides further evidence that it can be fruitfully thought of as an autonomous faculty, and reason not to interpret the difference between those who are willing and those who are not as a difference in the strengths of their respective desires. Are we to suppose that in performing the assigned 'willpower strengthening tasks' the subjects thereby made their desires weaker? All of their desires? Just the wayward ones which potentially interfere with their normative judgements and their intentions? How would that work? It seem that it would be much simpler to suppose that instead what had happened is that they have increased their effectiveness, and in the process their willingness, to resist temptations.

If we can get ourselves to take seriously the idea that there might be an autonomous psychological faculty that corresponds to the ordinary concept of willpower, it should be clear enough how this might bear on the situation of an addict who has formed the intention to quit. All things being equal, someone with more willpower should expect to find easier success at the implementation of this intention. But what implications does this have for the addict's responsibility? Does the addictive condition itself impair one's willpower? Are there external factors relevant for determining whether the failures of reactivity involved in addiction constitute excuses? I wish to conclude by considering these questions.

## 5.6 The Limits of Responsibility in Addiction

I hope to have illuminated three different kinds of volitional impairment in addiction: a self-imposed inability to recognize epistemic reasons; being subject to strong persistent appetite-like impulses that interfere with the ability to follow through with intentions; and having less willpower than one might wish to have. I hope to have made plausible that these are all

volitional phenomena, but it has surely been noticed that these are quite different sorts of impairments. The sense in which they are all impairments of the will is precisely, but perhaps unhelpfully, this: someone with an ideally functioning will would suffer from none of them. If my ability to respond to reasons and form the appropriate intentions were ideal, I would not be self-deceived (this is not therefore, a purely epistemic problem); if my ability to follow through with my intentions were ideal I would neither find myself subject to acquired A-impulses, nor would I find myself without the willpower required to overcome what wayward desires I did have (this is not therefore, a purely desiderative problem). But it doesn't follow straightaway from the fact that these impairments are bona fide impairments of the will that someone who suffers them is straightforwardly responsible for their condition. It is no part of the idea of the will that I have been working with that *all* volitional failures are themselves somehow willed or otherwise the result of voluntary actions and choices on the part of the agent.[32] I have argued that we should think that this is the case for addicts who are self-deceived, but the question of the extent to which addicts who suffer failures of reactivity are responsible for their conduct remains unsettled.

Wallace argues that the strength and persistence of the A-impulses might be enough to constitute a partial undermining of the agent's responsibility. I am largely in agreement. As I have noted at length, the volitional obstacles faced by someone who has a harmful acquired appetite with all the urgency of a natural one are indeed formidable. Wallace also acknowledges, however, that there is a complication: (in our terms) what about the *acquisition* of the appetite? There is also a further complication: The existence of an autonomous willpower capacity might be thought also to militate against the idea that A-impulses could partially excuse addicts if that capacity is itself something that one can take it upon oneself to cultivate. Let us take both complications in turn.

As far as the acquisition of the appetite goes, there is room for self-deception to be playing a role again. On my preferred view, a self-deceived agent is at least partially responsible for being self-deceived because there is a sense in which the resulting volitional impairment is the result of a voluntary action undertaken by the agent: acquiescing in the more comfortable, but externally defeated, belief. It seems quite plausible that the appetite for an addictive substance is acquired only over time by repeated ingestion. And to the extent to which self-deception may be implicated in the continued ingestion of the substance despite evidence of its ill-effects, the agent's responsibility for being self-deceived may carry over to this extended pattern of behaviour, and hence, partially, to the acquisition of the appetite. It is true that, by hypothesis, *self-deception* only comes about when the balance of reasons and evidence really does favour the negation of the thing that the agent has acquiesced in believing — otherwise it is merely a case of wishful thinking, or perhaps of lucky true belief.[33] But it is not required that the behaviour in fact becomes inconsistent with the agent's beliefs about what is disruptive or harmful for the agent to be blameworthy for believing what she in

[32] And even if they were, there would still be further questions that would need answering (such as whether those choices were themselves made under appropriate conditions etc.)

[33] See Chapter 3, fn. 11.

fact believes. And if her belief continues to be that everything is all right with her in a way which is insensitive to evidence, she may be blameworthy for continuing to perform the actions which led to the acquisition of the appetite.

As to the second complication: It is hopeful that with training people's willpower can become stronger. But we must ask: does the training required to strengthen one's willpower also require willpower? Presumably, it does. It may, if the results really do have a positive moral, be possible to, as it were, 'invest' a smaller amount of willpower now in the cultivation of one's overall willpower, making further investment and further cultivation easier and easier. But so long as the initial investment requires an exercise of willpower that is significant, it may simply be the case that someone in a deeply addicted condition will be unable to effectively get started. Being addicted, satisfying one's addictive impulses, and wreaking havoc on one's body is, perhaps first and foremost, *exhausting*.

What makes this even worse is that the socio-economic conditions that are associated with the highest rates of dangerous addiction are themselves conditions which make life exhausting for those that must endure them. The various facets of the human will that I have been emphasizing are not exercised in a vacuum. It is important that we not lose sight of the fact that human beings are endowed with a capacity for self-control, and even a capacity for increasing the effectiveness of this capacity for self-control, but the horizons within which those capacities are exercised can have a powerful limiting effect on how effective they can be.

Bruce Alexander has done some remarkable experiments on rats to test the effects of a positive stimulating environment on morphine addiction in rats. The basic so-called 'Rat-Park' experiments were designed to undercut the empirical support for the idea that mere exposure to drugs causes addiction which many scientists believed was provided by experiments on caged animals. Alexander quite rightly wondered whether the facts that caged animals would choose drug-laced water over pure water might have more to do with the fact that they housed in 'Skinner boxes' in isolation than with any 'irresistibility' of the drug. To test his hunch he built what he called 'Rat Park'. In his own words (Alexander 2010):

> A small group of colleagues at Simon Fraser University, including Robert Coambs, Patricia Hadaway, Barry Beyerstein, and myself undertook to test the conclusion about irresistibly of addicting drugs that had been reached from the earlier rat studies. We compared the drug intake of rats housed in a reasonably normal environment 24 hours a day with rats kept in isolation in the solitary confinement cages that were standard in those days. This required building a great big plywood box on the floor of our laboratory, filling it with things that rats like, such as platforms for climbing, tin cans for hiding in, wood chips for strewing around, and running wheels for exercise. Naturally we included lots of rats of both sexes, and naturally the place soon was teeming with babies. The rats loved it and we loved it too, so we called it "Rat Park".

What Alexander and his colleagues found was that compared to rats raised in captivity

and who spend their lives in cages, the rats in Rat Park took morphine at much lower rates, despite the fact that it was just as easily obtained. Even more remarkable, Alexander found that even rats who were bred and spent 57 days in isolation choosing morphine-laced water over pure water started dispreferring the morphine water when they were transfered to Rat Park.

Rats are not possessed of the same volitional capacities as human beings, but even without a strong capacity for self-regulation they come to disprefer a life of drug-taking when they are relocated to somewhere intentionally designed to promote their flourishing. It is thus eminently plausible that many limitations on the human will are imposed from without. A life of poverty, violence, and exclusion may cause someone to be highly tempted by the prospect of escape into drugs but could also make one more prone to self-deception, and constitute an external limitation on one's willpower. It could also seriously limit one's ability to cultivate the further willpower that might be needed to extricate oneself from the addictive condition. Any of these factors could be partially mitigating. External impediments are rarely decisively coercive, but it would certainly be inhumane to ignore their force. What is wrong, therefore, with the liberal conception of addiction and addictive behaviour is that it supposes that there is no space to talk about volition and choice in addiction at all. That is false, even though the external obstacles faced by addicts — strong appetites showing up as brain lesions, diminished willpower, dismal socio-economic prospects etc. — are considerable and deserve to be acknowledged. However, the conservative view is equally wrong insofar as it supposes that, from the fact that addiction and addictive behaviour *involve* choice, it follows that the agent is fully responsible for her conduct. Forming intentions and trying one's best to follow through with them are acts of volition, but they are not thereby radically, or limitlessly free. Human beings are fundamentally reasons-responsive creatures, not just organisms pushed around inside fields of force. But, for all that, we are also organisms pushed around inside fields of force. The capacities we have for regulating our conduct make us morally responsible agents, but the exercise of those capacities can be made difficult enough that it would be unreasonable to expect someone to succeed.

These conclusions may strike many as unsatisfactory. I have not simply and decisively answered the question 'Are addicts morally responsible for their conduct?'. But I don't think any clear or snappy answer to that question is plausible. In the case of any individual addict, the answer will be 'it depends'. It depends on what condition she is in and whether she is aware of that condition or not. If she isn't aware, it depends on why this is so. It depends on how strongly entrenched her acquired appetite has become and how much willpower she has to resist. If she lacks the willpower, it depends on why *that* is so. I do not see any way of eliminating these complexities from philosophical thinking about addiction. The best we can hope for is to cast the phenomenon under a theory of moral responsibility which is capable of handling these nuances.

# Bibliography

Alexander, Bruce. 2008. *The Globalisation of Addiction: A Study in Poverty of the Spirit.* Oxford University Press.

———. 2010. "Addiction: The View from Rat Park." `http://www.brucekalexander.com/articles-speeches/rat-park/148-addiction-the-view-from-rat-park`.

Alexander, Bruce, Robert Coambs, and Patricia Hathaway. 1978. "The Effects of Housing and Gender on Morphine Self-Administration in Rats." *Psychopharmacology* 58 (2): 175–179.

Aristotle. 1999. *Nichomachean Ethics.* Edited by Terence Irwin. Hackett Publishing: Indiana.

Arpaly, Nomy. 2001. *Unpricnipled Virtue: An Inquiry into Moral Agency.* Oxford University Press.

Barnes, Annette. 1997. *Seeing Through Self-Deception.* Cambridge University Press.

Baumeister, Roy. 2012. "Self-Control: The Moral Muscle." *The Psychologist* 25 (2): 112–115.

Bayne, Tim, and Elizabeth Pacherie. 2004. "Bottom-up or Top-down? Campbell's Rationialist Account of Monothematic Delusions." *Philosophy, Psychiatry, and Psychology* 11:1–11.

Borge, Steffen. 2003. "The Myth of Self-Deception." *Southern Journal of Philosophy* 41 (1): 1–28.

Bortolotti, Lisa. 2015. "The Epistemic Innocence of Motivated Delusions." *Consciousness and Cognition* 33:490–499.

———. 2016. "Epistemic Benefits of Elaborated and Systematized Delusions in Schizophrenia." *British Journal of Philosophy of Science* 67:879–900.

Boyd, Richard. 1999. "Homeostatis, Species, and Higher Taxa." In *Species: New Interdisciplinary Essays,* edited by Robert A. Wilson. MIT Press.

Bruner, Jerome. 1960. *The Process of Education.* Harvard University Press.

Burroughs, William. 1977. *Junky.* Penguin Books.

Butler, Peter. 2000. "Reverse Othello Syndrome Subsequent to Traumatic Brain Injury." *Psychiatry* 63:85–92.

Casey, B. J., Leah H. Somerville, Ian H. Gotlib, Ozlem Ayduk, Nicholas T. Franklin, Mark K. Askren, John Jonides, et al. 2011. "Behavioral and Neural Correlates of Delay of Gratification 40 Years Later." *Proceedings of the National Academy of Sciences* 108 (36): 14998–15003.

Clifford, William. 1999. "The Ethics of Belief." In *The Ethics of Belief and other Essays,* edited by Timothy J. Madigan, 70–96. Prometheus Books.

Coltheart, Max. 2005. "Conscious Experience and Delusional Belief." *Philosophy, Psychiatry, & Psychology* 12:152–157.

Currie, Gregory. 2000. "Imagination, Delusions, and Hallucinations." In *Mind and Language,* edited by Max Coltheart and Martin Davies, 168–183. Blackwell.

Darwall, Stephen. 1988. "Self-Deception, Autonomy, and Moral Constitution." In *Perspectives on Self-Deception,* edited by Brian Mclaughlin and Amelie Oksenberg Rorty, 407–430. University of California Press.

Davidson, Donald. 2004a. "Deception and Division." In *Problems of Rationality,* 200–210. Oxford University Press.

———. 2004b. "Paradoxes of Irrationality." In *Problems of Rationality,* 169–188. Oxford University Press.

———. 2004c. "Representation and Interpretation." In *Problems of Rationality,* 83–100. Oxford University Press.

Davies, Martin. 2010. "Delusion and Motivationally Biased Belief: Self-Deception in the Two-Factor Framework." In *Delusion and Self-deception: Affective and Motivational Influences on Belief Formation,* edited by Tim Bayne and Jose Fernández, 71–86. Psychology Press.

Davies, Martin, Max Coltheart, Robyn Langdon, and Nora Breen. 2001. "Monothematic Delusions: Toward a Two-Factor Account." *Philosophy, Psychiatry, & Psychology* 8 (1/2): 133–158.

D'Cruz, Jason. In preparation. "Self-Deception as Unwitting Pretense."

Dennett, Daniel. 1981. "True Believers: The Intentional Strategy and Why it Works." In *Scientific Explanantions,* edited by A. F. Heath. Oxford University Press.

Doggett, Tyler. 2012. "Some Questions for Tamar Gendler." *Analysis* 21 (1): 231–258.

Doris, John. 2015. *Talking to Ourselves: Reflection, Ignorance, and Agency.* Oxford University Press.

Fingarette, Herbert. 1969. *Self-Deception.* University of California Press.

———. 1988. *Heavy Drinking: The Myth of Alcoholism as a Disease.* University of California Press.

Fischer, John Martin, and Michael Ravizza. 1998. *Responsibility and Control: An Essay on Moral Resposibility.* Cambridge Univerity Press.

Foucault, Michel. 1969. *L'archéologie du savoir.* (The Archaeology of Knowledge). Translated by Allan Sheridan. Harper & Row.

Frankish, Keith, and Jonathan St. B. T. Evans. 2009. "The Duality of Mind: An Historical Perspective." In *In Two Minds: Dual Processes and Beyond,* 1–29. Oxford University Press.

Freud, Sigmund. 1963. "The Unconscious." In *General Psychological Theory: Papers on Metapsychology,* edited by Philip Rieff, 116–150. Collier Books.

Gendler, Tamar. 2007. "Self-Deception as Pretense." *Philosophical Perspectives* 21 (1): 231–258.

Goldstein, Avram. 1997. "Addiction and the Brain." `http://www.aatod.org/OLD_SITE/print_version/print_1998-3.html`.

Gomez, Rebecca. 1997. "Transfer and Complexity in Artificial Grammar Learning." *Cognitive Psychology* 44:154–207.

Heyman, Gene. 2009. *Addiction: A Disorder of Choice.* Harvard University Press.

Holton, Richard. 2009. *Willing, Wanting, Waiting.* Oxford University Press.

Hume, David. 2000/1738. *A Treatise of Human Nature.* Edited by David Fate Norton and Mary J. Norton. Oxford University Press.

Islam, Lucrezia, Sylvie Piacentini, Silvio Scarone, and Orsola Gambini. 2015. "Capgras Delusion for Animals and Inanimate Objects in Parkinson's Disease: A Case Report." *BMC Psychiatry* 15:73.

Jaspers, Karl. 2007. "Causal and 'Understandable': Relationships Between Events and Psychosis in Dementia Praecox (Schizophrenia)." In *Anthology of German Psychatric Texts,* edited by Henning Sass, 174–279. Blackwell Publishing: Oxford.

Johnston, Mark. 1988. "Self-Deception and the Nature of Mind." In *Perspectives on Self-Deception,* edited by Brian Mclaughlin and Amelie Oksenberg Rorty. University of California Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Farrar, Straus, & Giroux.

Kendler, Kenneth S., and John Campbell. 2014. "Expanding the Domain of the Understandable in Psychiatric Illness: An Updating of the Jasperian Framework for Explanation and Understanding." *Psychological Medicine* 44 (1): 1–7.

Kim, Jaegwon. 2003. "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation." *Philosophy and Phenomenological Research* 67 (1): 151–176.

Korsgaard, Christine. 2009. *Self-Constitution: Agency, Indentity, and Integrity.* Oxford University Press.

Kripke, Saul. 1980. *Naming and Necessity.* Basil Blackwell: Oxford.

List, Christian, and Peter Menzies. 2007. "Non-Reductive Physicalism and the Limits of the Exclusion Principle." *LSE Working Papers.*

Locke, John. 1975/1690. *An Essay Concerning Human Understanding.* Edited by Peter H. Nidditch. Oxford University Press.

Lockie, Robert. 2003. "Depth Psychology and Self-Deception." *Philosophical Psychology* 16 (1): 127–148.

Maher, Brendan. 1974. "Delusional Thinking and Perceptual Disorder." *Journal of Individual Psychology* 30:98–113.

Mele, Alfred. 1997. "Real Self-Deception." *Behavioral and Brain Sciences* 20:91–136.

———. 2006. "Self-Deception and Delusions." *European Journal of Analytic Philosophy* 2 (1): 109–124.

Mischel, Walter. 1996. "From Good Intentions to Willpower." In *The Psycolgoy of Action,* edited by Perter Gollwitzer and John Bargh. Guilford Press: New York.

Mischel, Walter, Ozlen Ayduk, Marc G. Berman, B.J. Casey, Ian H. Gotlib, John Jonides, Ethan Kross, et al. 2011. "'Willpower' Over the Life Span: Decomposing Self-Regulation." *Social Cognitive and Affective Neuroscience* 6(2):252–256.

Muraven, Mark. 2010. "Praciticing Self-Control Lowers Risk of Smoking Lapse." *Psychology of Addictive Behaviors* 24 (3): 446–452.

Muraven, Mark, Roy Baumeister, and Dianne Tice. 1999. "Longitudinal Improvement of Self-Regulation Through Practice: Building Self-Control Strength Through Repeated Exercise." *The Journal of Social Psychology* 131 (4): 446–457.

Nisbett, Richard, and Timothy Wilson. 1977. "Telling More than We Can Know." *Psychological Review* 84 (3): 231–259.

Oaten, Megan, and Ken Chang. 2006. "Longitudinal Gains in Self-Regulation From Regular Physical Exercise." *British Journal of Health Psychology* 11 (4): 717–733.

Pears, David. 1984. *Motivted Irrationality.* Oxford University Press.

Pelham, Brett W., Matthew C. Mirenberg, and John T. Jones. 2002. "Why Susie Sell Seashells by the Seashore: Implicit Egotism and Moajor Life Decisions." *Journal of Personality and Social Psychology* 82 (4): 469–487.

Pianezza, M., E. Sellers, and R. Tyndale. 1998. "Nicotine Metabolism Defect Reduces Smoking." *Nature* 393:750.

Plato. 1997a. "Protagoras." In *Plato: Complete Works,* edited by John Cooper, translated by Stanley Lombard and Karen Bell, 746–791. Hackett Publishing: Indiana.

———. 1997b. "The Republic." In *Plato: Complete Works,* edited by John Cooper, translated by G. M. A. Grube, 971–1223. Hackett Publishing: Indiana.

Ramachandran, Vilayanur S. 1996. "The Evolutionary Biology of Self-Deception, Laughing, Dreaming, and Depression: Some Clues from Anosognosia." *Medical Hypotheses* 47 (5): 602–632.

Samuels, Richard. 2009. "The Magical Number Two, Plus or Minus: Dual-process Theory as a Theory of Cognitive Kinds." In *In Two Minds: Dual–processes and Beyond,* edited by Jonathan St. B. T. Evans and Keith Frankish, 129–146. Oxford University Press.

Sartre, Jean-Paul. 1966. *Being and Nothingness.* Edited by Hazel Barnes. Citadel Press: New York.

Scanlon, Thomas. 1998. *What we Owe to Each Other.* Harvard University Press.

Schneider, Walter, and Richard Shiffrin. 1977a. "Controlled and Automatic Human Information Processing: I Detection Search and Attention." *Psychological Review* 84 (1): 1–66.

———. 1977b. "Controlled and Automatic Human Information Processing: II Perceptual Learning Automatic Attending, and a General Theory." *Psychological Review* 84 (2): 127–190.

Shoemaker, David. 2011. "Attributability, Answerability, Accountability." *Ethics* 121 (3): 602–632.

———. 2015. "Review of John Doris' *Talking to Ourselves: Reflection, Ignorance, and Agency.*" *Notre Dame Philosophical Reviews.* `http://ndpr.nd.edu/news/talking-to-our-selves-reflection-ignorance-and-agency/`.

Sims, Andrew. 2003. *Symptoms in the Mind: An Introduction to Desriptive Psychopathology.* 3rd Edition. Saunders: Elsevier Science.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind* 96(381):36–61.

Smtih, Holly. 1983. "Culpable Ignorance." *The Philosophical Review* 92 (4): 543–571.

Strawson, Peter. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48:1–25.

Summers, Jesse. 2011. "Addiciton as Compulsive Misvaluation." PhD diss., UCLA.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (1): 207–233.

Volkow, Nora. 2015. "Addiction: A Disease of Free Will." 168th Meeting of The American Psychiatric Association.

Wallace, R. Jay. 1990. "How to Argue about Practical Reason." *Mind* 99 (395): 355–385.

———. 1994. *Responsibility and the Moral Sentiments.* Harvard University Press.

———. 1999a. "Addiction as a Defect of the Will: Some Philosophical Reflections." *Law and Philosophy* 18(6):621–654.

———. 1999b. "Three Conceptions of Rational Agency." *Ethical Theory and Moral Practice* 2 (3): 217–242.

Wason, Peter, and Jonathan St. B. T. Evans. 1975. "Dual Processes in Reasoning?" *Cognition* 3:141–154.

Watson, Gary. 2004a. "Disordered Appetites: Addiction, Compulsion, and Dependence." In *Agency and Answerability: Selected Essays.* Oxford University Press.

———. 2004b. "Two Faces of Responsibility." In *Agency and Answerability,* 260–288. Oxford University Press.

Weedon, Chris. 1987. *Feminist Practice and Poststructuralist Theory.* Blackwell Publishing.

Williams, Bernard. 1972. "Deciding to Believe." In *Problems of the Self,* 136–151. Cambridge University Press.

Yablo, Stephen. 1992. "Mental Causation." *Philosophical Review* 101 (2): 245–280.