Social Constraints on Human Agency

Andreas Paraskevaides

PhD

The University of Edinburgh

2010

Abstract

In this thesis, I present a view according to which folk psychology is not only used for predictive and explanatory purposes but also as a normative tool. I take it that this view, which I delineate in chapter 1, can help us account for different aspects of human agency and with solving a variety of puzzles that are associated with developing such an account. My goal is to examine what it means to act as an agent in a human society and the way in which the nature of our agency is also shaped by the normative constraints inherent in the common understanding of agency that we share with other agents. As I intend to demonstrate, we can make significant headway in explaining the nature of our capacity to express ourselves authoritatively in our actions in a self-knowing and self-controlled manner if we place this capacity in the context of our social interactions, which depend on a constant exchange of reasons in support of our actions. My main objective is to develop a promising account of human agency within a folk-psychological setting by mainly focusing on perspectives from the philosophy of action and mind, while still respecting more empirically oriented viewpoints from areas such as cognitive science and neuroscience.

Chapter 2 mainly deals with the nature of self-knowledge and with our capacity to express this knowledge in our actions. I argue that our self-knowledge is constituted by the normative judgments we make and that we use these judgments to regulate our behaviour in accordance to our folk-psychological understanding of agency. We are motivated to act as such because of our motive to understand ourselves, which has developed through our training as self-knowing agents in a folk-psychological framework. Chapter 3 explores the idea that we develop a self-concept which enables us to act in a self-regulating manner. I distinguish self-organization from self-regulation and argue that we are self-regulating in our exercises of agency because we have developed a self-concept that we can express in our actions. What makes us distinct from other self-regulating systems, however, is that we can also recognize and respond to the fact that being such systems brings us under certain normative constraints and that we have to interact with others who are similarly constrained. Chapter 4 is mainly concerned with placing empirical evidence which illustrate the limits of our conscious awareness and control in the context of our account of agency as a complex, emergent social phenomenon. Finally, chapter 5 deals with the way in which agentive breakdowns such as self-deceptive inauthenticity fit with this account.

Declaration

Pursuant to The University of Edinburgh's Postgraduate Research Assessment Regulations, (section 2.5) I hereby declare that the thesis has been composed by me, that the work is my own, and that it has not been submitted for any other degree or professional qualification.

Andreas Paraskevaides, 21/04/10

Table of Contents

Chapter 1-Introduction

Folk Psychology as the Playground of Agency

Introduction

The main question I intend to examine in this thesis is this: Can we provide an account of human agency that strikes a balance between our sense of being in control of our actions and our nature as complex, physically constrained organisms? When I talk about our sense of being in control of our actions, I have in mind various assumptions that seem to intuitively fit our conception of agency. We have the capacity to do more than just passively react to changes in our environment. We can deliberate and assess our circumstances, make decisions, choices and plans, and act as we judge best. We have a special kind of authorship over our behaviour, because we can engage in actions that express our unique point of view on the world, our first-person perspective on our circumstances. We have reasons for acting in the ways we do, and can invoke these reasons in explaining our behaviour. Driven by the assumption that we have the final say on our actions, we routinely hold each other responsible for them, offering criticism and praise as we see fit. The actions we engage in as agents are self-determined; they express ourselves and our active contribution in our behaviour. They are self-knowing; we know what we're doing and why, because we know ourselves and can explain what was in our mind when deciding to act in a certain way. They are self-controlled; they manifest a kind of unity and purpose that is the mark of active control. The agent is more than just a product of his environment, or an outcome of a series of events, or a reaction to a set of circumstances. The agent is also a creator, a cause and a controller.

Taking a closer look at these initial assumptions, we immediately encounter pressing issues that need to be clarified. What is it that we express exactly, when we express ourselves in our actions? What is referred to as the "self" in terms such as self-control, self-knowledge and self-determination? What kind of knowledge constitutes self-knowledge? Where does the unity and purpose displayed in our actions come from? What kind of control is exercised in expressions of agency? What constitutes our reasons for acting? What justifies our ascriptions of responsibility? How can we reconcile these assumptions with the fact that we are also parts of a natural order, empirically constrained, influenced by a variety of factors such as our personal history, our complex physical nature and our present circumstances? How can our actions be exercises of our own active agency, the product of authoritative control on our part, when they are also situated within a complex chain of physical causes and effects? Can we ever be self-deceived about our reasons for acting?

In developing an account of agency that can address such issues, I will operate from a physicalist's perspective and avoid any answers that assume that we must somehow override our physical nature and the chain of causes and effects that we are part of in order to act as agents. I will also argue that our initial assumptions are not fundamentally misguided, since we do have the capacity to express ourselves in our actions as active, self-controlled individuals that know their own minds and actions. I will show that we can develop an account of agency that respects both our status as authoritative agents and our place in the natural order. The key to developing such an account is the fact that when we act as agents, we don't act alone. We are social creatures that collaborate in maintaining a framework wherein we respond to each other as self-knowing agents. In this chapter, I will examine the debate on how to properly understand what we do when we treat each other as self-knowing agents and argue that there is a specific conception of our social nature and our collaborative practices that can serve as the foundation for a promising account of agency.

The traditional conception of folk psychology

In the course of our interaction with each other as parts of a complex, developing society, we regularly engage in interpretations of each other's behaviour. When examining how we behave, we tend to treat ourselves and others as having certain intentional states and characteristics (such as certain beliefs, desires, hopes, fears and particular inclinations) which are displayed in this behaviour. This kind of treatment enables us to explain our own and others' behaviour and to predict the forms that future behaviour will take. These explanations and predictions are made with varying accuracy, depending on the interpretations they are based on.

It's also worth noting that a common assumption in such collaborative practices is that for an intentional system such as a human being to be considered an agent, it must be in control of its behaviour in a way that fits with the intentional characterizations attributed to it[1]. It must, in other words, act in accordance with what are commonly perceived as its beliefs, desires and other intentional characteristics. Hence, following this common assumption, the practice of interpretation also leads to a practice of criticism when the intentional behaviour interpreted frequently fails to conform to its most common interpretation. Systems acting under intentional descriptions can then be held responsible not only for their actions but also for failing to act in ways coherent with the prevailing intentional interpretation of their behaviour.

In the philosophical and psychological literature that focuses on these practices they have all been lumped under one common heading: folk psychology. The origins, exact functions and overall purpose of folk psychology are hotly contested in the relevant discussions, but traditionally, these discussions all share the assumption that folk psychology is the practice that human beings engage in when interpreting, explaining and predicting each other's behaviour. The rest is up for grabs.

The two prevalent theories concerning folk psychology are the theory-theory and the simulation theory[2]. Briefly, the proponents of theory-theory argue that when humans engage in folk psychological ascriptions of mental states, they utilize an underlying theory having to do with the nature of intentional states such as beliefs and desires and with the way these are manifested in behaviour. This underlying theory is also based on generalizations of certain observed behavioural patterns, as well as on general knowledge of the causes of mental states such as beliefs and desires, the ways these states can relate to one another and the effects these states can

[1] I use the term "intentional system" as it is used by Daniel C. Dennett in his theory of intentionality (see especially Dennett, 1981/2008, "True Believers: The Intentional Strategy and Why it Works", in W.G. Lycan and J.J. Prinz (eds.) *Mind and Cognition: An Anthology* (3rd edition), Blackwell Publishing Ltd, pp. 323-336 and Dennett, 1991/2008, "Real Patterns", *ibid*, pp. 351-366. According to Dennett, something is an intentional system if its behaviour can be explained and predicted by using what he calls "the intentional stance". The users of the intentional stance use their folk-psychological understanding of intentionality to interpret the behaviour of objects they encounter (which can also be their own behaviour). If this behaviour conforms to this understanding and can be reliably predicted in accordance to it, then it is the behaviour of an intentional system. I will discuss Dennett's views and their relevance to my own account in much more detail in later chapters (see especially Chapters 3 and 4).

[2] For comprehensive overviews of simulation and theory-theories which have been essential in helping me structure my presentation of these differing viewpoints, see Tony Stone and Martin Davies, 1996, "The Mental Simulation Debate: A Progress Report", in P. Carruthers and P.K. Smith (eds), *Theories of Theories of Mind,* Cambridge: Cambridge University Press, pp.119-137, Davies and Stone, 2000, "Simulation Theory", Entry for *Routledge Encyclopedia of Philosophy Online* and Shaun Nichols, "Folk Psychology", in *Encyclopaedia of Cognitive Science* London: Nature Publishing Group, pp. 134-140.

have on behaviour. For example, one such generalization might be that when someone believes that a certain person is extremely unpleasant to talk to, he will generally tend to avoid that person in most circumstances. And vice versa, when someone tends to avoid a person in most circumstances then one of the beliefs that might be plausibly attributed to him is the belief that this person is extremely unpleasant to talk to. In such a way an underlying psychological theory based on similar generalizations can be used in practices of interpretation, prediction and explanation. Furthermore, according to theory-theorists this psychological theory is not explicitly formulated by its users but it nevertheless underlies the practices of intentional interpretation, explanation and prediction of behaviour they engage in[3].

A different explanation of folk psychological ascription is offered by the proponents of the simulation theory. According to these theorists, what is crucial for one's attributions of intentional states to one's self and others is not the use of an implicit psychological theory based on generalizations of behavioural patterns, but the use of simulation and imagination in action. What these theories emphasize is that folk psychological attributions depend on the imaginative recreation of the same processes that lead to action in the case of the behaviour that is to be explained. According to the simulation theorists, by recreating the behaviour of an observed intentional system one can be led to the recreation of the intentional states that led to such behaviour. And by simulating an interaction of certain intentional states and characteristics with certain circumstances one can be led to simulating the effects of this interaction and the behaviour that it would actually produce. For example, the desire to avoid someone can be arrived at by focusing on that person and simulating the belief that he is extremely unpleasant to talk to. The simulation of the psychological characteristics and the circumstances of an intentional system can lead then to the prediction and explanation of its behaviour, assuming that the simulator possesses the capacity to perform a simulation of this sort. In the case of human

[3] For different accounts in favour of the use of theorizing in our folk-psychological practices, see P. Carruthers, 1996, "Simulation and Self-Knowledge: A Defence of Theory-Theory", in P. Carruthers and P.K. Smith (eds.) *Theories of Theories of Mind*, Cambridge University Press, pp.22-39, S.P. Stich and S. Nichols, 1992, "Folk Psychology: Simulation or Tacit Theory?", *Mind and Language* 7(1), 1992, pp. 35-71, Stich and Nichols, 1995, "Second Thoughts on Simulation", in T. Stone And M. Davies (eds.) *Mental Simulation: Evaluations and Applications,* Oxford: Blackwell Publishers, pp. 87-108, A. Gopnik and H.Wellman, 1992, "Why the Child's Theory of Mind Really is a Theory", *Mind and Language* 7, pp. 145-171, A.Gopnik, Gopnik, 1993, "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality", *Behavioral and Brain Sciences* 16, pp.1-14

beings, as the simulation theorists argue, they are similar enough in their functions and capacities to be able to successfully explain, predict and interpret each other's behaviour by the use of imagination and simulation[4].

There are a lot of subtleties in the opposition of these two theories in the literature that have not been covered with this brief exposition of their main points. For example, an important argument advanced by the theory-theorists against simulation theories is that theories based on simulation cannot readily explain mistakes made in attributing intentional states or in explaining, predicting or interpreting intentional behaviour. The theory-theorists can explain such mistakes by alluding to the fact that in such cases a faulty psychological theory of intentional behaviour underlies these false attributions. A relevant example can be found in the literature focusing on the development of children's folk psychological capacities. In what has been termed the "false belief" task, three year olds observe a puppet placing a candy in a box and leaving the room[5]. While the children still observe the room, the candy is moved into another box and the puppet is brought back into the room. When the children are asked to predict where the puppet believes the candy can be found, they say that it believes the candy is in the second box. The theory-theorists explain this mistake by pointing out the fact that these children seem to have a different understanding of the causes and effects of intentional behaviour. Specifically, they seem to have as a background belief that whatever is actually occurring in the world is accurately represented by the mental states of intentional systems.

Simulation theorists have offered some replies to this argument and against adopting an account that focuses solely on the possession of an implicit psychological theory that underlies folk-psychological ascription. In the case of

[4] For different accounts in favour of the use of simulation in our folk-psychological practices, see A. Goldman, 1989, "Interpretation Psychologized", *Mind and Language* 4, pp.161-185, Goldman, 1992, "In Defense of the Simulation Theory", *Mind and Language* 7(1), pp. 104-119, Goldman, 1993, "The Psychology of Folk Psychology", *Behavioral and Brain Sciences* 16, pp. 15-28, R. Gordon, 1986, "Folk Psychology as Simulation", *Mind and Language* 1, pp. 158-171, Gordon, 1992, "The Simulation Theory: Objections and Misconceptions, *Mind and Language* 17, pp.11-34, Gordon, 1995, "Simulation Without Introspection or Inference From Me to You", in T. Stone and M. Davies (eds.) *Mental Simulation: Evaluations and Applications*, Oxford: Blackwell Publishers, pp. 53-67 and J. Heal (1998), "Co-cognition and Offline Simulation: Two Ways of Understanding the Simulation Approach", *Mind and Language* 13(4), pp. 477-498.

[5] See H. Wimmer and J. Perner, 1983, "Beliefs about Beliefs: Representation and the Containing Function of Wrong Beliefs in Young Children's Understanding of Deception", *Cognition* 13, pp. 103-128.

accounting for mistakes in folk psychological practices, simulation theorists have argued that perhaps the source of these mistakes in not a faulty psychological theory but a physical factor that cannot be accounted for in simulation. An example offered is that of an intentional system's suffering from fatigue or from the effects of a drug. Even if simulation of the psychological characteristics of this system is perfect, a mistake might still be made due to these unforeseen physical factors. Subsequently, simulation theorists have argued that it is perhaps a combination of simulation and of knowledge of such physical factors that informs the simulation that is used in successful folk- psychological ascription. Another argument that has been offered in the simulation side of the debate is that engaging in folk psychological practices by simulation is much simpler to account for than having a theory formed from knowledge of generalizations in behavioural patterns. The argument in the case of human folk- psychological ascription seems to be that since human beings are similar in respect to the psychological and physical factors that influence their behaviour, all they would have to do when engaging in folk-psychological practices would be to imagine themselves in the same situations as facing the subjects whose behaviour they wish to examine. This seems to be less complicated than the requirement that they would have to first form a theory based on generalizations and then apply this theory in each individual case.

In any case, the debate between simulation theorists and theory-theorists remains unresolved. The possibility of hybrid theories has been suggested[6], but what is of interest for my current purposes is that both these theories seem to have some common views on folk-psychological practices. It seems that for the theorists engaging in this debate, the practice of folk psychology has as its main purpose the explanation, prediction and interpretation of intentional behaviour. And this purpose is conducive to the stable function of human societies where everyone has to successfully communicate with each other so as to engage in all the social cooperative practices required for this function. But there are alternative conceptions

[6] For example, Stone and Davies write that "[t]the mental simulation debate has reached a stage in which there is considerable agreement about the need to develop hybrid theories-theories that postulate both theory and simulation, and then spell out the way in which these two components interact" (Stone and Davies, 1996, p.136), while Nichols notes that "[a]lthough it's likely that the theory theory explains part of the capacity for mindreading, it's also likely the theory theory cannot provide anything remotely like a complete account for the capacity for mindreading, [which] also plausibly depends on simulation-like processes[.]" (Nichols, 2002, Conclusion).

of folk psychology, according to which there is more to it than the predictive and explanatory practices emphasized in the simulation and theory-theory approaches[7]. As will become evident in the following discussion of one such departure from the traditional way of seeing folk psychology, views focusing on just the predictive and explanatory function of folk psychology seem to leave out one of its crucial aspects: its normative function as the collaborative enterprise that is used in training human beings to think and act as responsible agents. More specifically, Victoria McGeer's work on self-knowledge and the normative role of folk psychology[8] can provide the template for exploring such an alternative conception.

The alternative conception of folk psychology

McGeer develops an account of intentional states in which they are dispositional, in that they consist of the agent's dispositions towards the feelings, thoughts and actions that express these states[9]. These feelings, thoughts and actions can be integrated as a coherent whole expressing the agent's intentional states because of the agent's active contribution to his actions. The agent is motivated to express his intentional states in the ways he does because of the claims he has made (privately and publicly) expressing his various tendencies and pro-attitudes to act in certain ways. These self-attributions function as normative commitments for the agent, as McGeer puts it, because they motivate him to bring his behaviour in line with the way he sees himself, and the way the agent sees himself, his self-conception, is based on his self-attributions:

[7] See, for example, S. Gallagher, 2001, "The Practice of Mind: Theory, Simulation, or Primary Interaction?", *Journal of Consciousness Studies* 8, pp. 83-108, D. Hutto, 2004, "The Limits of Spectatorial Folk-Psychology", *Mind and Language* 19(5), pp. 548-573, Hutto, 2007, "Folk-Psychology Without Theory or Simulation", in D. Hutto, M. Ratcliffe (eds.) *Folk Psychology Re-Assessed*, pp. 115-135 and T. W. Zawidski , 2008, "The Function of Folk Psychology: Mind Reading or Mind Shaping?", *Philosophical Explorations* 11(3), pp.193-210. These authors hold different views but they all agree that there is more to our folk-psychological practices than their predictive and explanatory aspects.

[8] See Victoria Mc Geer and Philip Pettit, 2002, "The Self-Regulating Mind", *Language and Communication,* Vol.22, no.3, pp.281-299, Mc Geer, 1996, "Is Self-Knowledge an Empirical Problem? Renegotiating the Space of Philosophical Explanation", *Journal of Philosophy* 93*,* pp. 483-515, McGeer, 2001, "Psycho-Practice, Psycho-Theory and the Contrastive Case of Autism: How Practices of Mind Become Second-Nature", *Journal of Consciousness Studies* 8(5-7), pp. 109-132, McGeer, 2007a, "The Moral Development of First-Person Authority, *European Journal of Philosophy* 16(1), pp.81-108 and McGeer, 2007b, "The Regulative Dimension of Folk Psychology", in D.Hutto and M.Ratcliffe (eds.) *Folk Psychology Re-Assessed,* Dordrecht: Springer, pp. 138-156.

[9] See, e.g. McGeer, 1996, pp. 506-508 and McGeer 2007a, p.90: "[B]eliefs and desires are complex dispositions to think, speak, feel and otherwise operate in various mental and physical ways."

> "Put simply, we are able to *ensure* a fit between the psychological profile we create of ourselves in first-person utterances and the acts our self-attributed intentional states are meant to predict and explain simply by adjusting our actions in appropriate ways." (Mc Geer, 1996, p. 507)

The agent in this account does not have an intimate knowledge of his own intentional states by somehow detecting certain phenomenal qualities that he can accurately report as signifying the occurrence of these states within him. He does not either directly experience these states or infer them from his own behaviour. Instead, he has such intimate knowledge because he has undertaken certain normative commitments to behave in ways expressing the kind of person he describes himself as being. The agent's own first person perspective is hence crucial to his expression of agency in his actions, since he can reject or reinforce the normative commitments he undertakes and thus influence his behaviour in a different way than someone who examines this agent's intentional states from a third person perspective. From the agent's own point of view, his intentional states are not simply independent objects of perception that occur within him and which he reports as a passive recipient of their effects. Instead these states are expressed as such in his behaviour (in his thoughts, feelings and actions) partly because of his own active participation in expressing them as such.

In McGeer's view, if an agent's behaviour becomes particularly discordant with the way he sees himself, then he can either revise his self-attributions or attempt to behave in ways that are more expressive of these self-attributions. In this case the aforementioned practice of criticizing an intentional system's behaviour will also be viewed as a practice of intentional correction. The agent who is criticized for a failure to act as is rationally expected of him (this form of evaluation may also have as a source the agent himself, when he recognizes that his behaviour does not fit with his self-conception) is motivated to adjust his behaviour in relevant ways, in order to maintain his status as a rational agent with whom other rational agents can cooperate in social contexts. If an agent regularly fails to act in such a way, then he will plausibly lose his status as such an agent who can be held responsible, by himself and others, for his actions. And what is the consequence of losing one's status as a rational agent?

> "At the extreme, the consequences of a general loss of authority, for good or bad reasons, are dire indeed. They involve various sorts of disenfranchisement-social, political, economic, legal, moral, and so on. An individual so treated becomes a patient rather than an agent, one whose behaviour is first figuratively and then literally taken out of her control." (McGeer, 1996, p. 509)

This emphasis on the role of the agent in living up to self-attributions that express his intentional states has distinct implications for the way we should understand folk-psychological practices. In her (2007b), McGeer expands on the differences between her view and the traditional way in which folk-psychological practices have been understood. In focusing on folk psychology used as a predictive and explanatory tool, she argues that "we overlook the way folk psychology operates as a *regulative* practice, moulding the way individuals act, think and operate so that they become well-behaved folk-psychological agents: agents that can be well-predicted and explained using both the concepts and the rationalizing narrative structures of folk psychology." (McGeer, 2007b, p. 139). In this view, our predictive and interpretative successes in understanding one another from a folk-psychological perspective depend on our capacity to act as agents that can be understood from this kind of perspective. This makes the reason this perspective has persisted through our dealings with one another less mysterious than it is in approaches that try to account for this fact by looking at how successful we are in interpreting and predicting one another's behaviour.

In defence of the alternative conception

Having set out the two viewpoints concerning the attribution of intentional states and the role of folk psychology, a question might be why we should prefer the alternative conception of folk psychology which focuses on the role of the agent in action and on the normative role of folk-psychological attributions, over the traditional one focusing only on the explanation and prediction of intentional action through the use of folk psychology. After all, perhaps an interpretation of intentional action that leads to explanation and prediction might be all that is needed to account for how it is that folk psychological practices are conducive to the stable function of society. Human beings are able to successfully predict and explain each other's behaviour using either simulation or an implicit theory of the causes and consequences of intentional action (while using the intentional idiom consisting of concepts such as beliefs and desires), or even a combination of the two. And explanation and prediction of intentional behaviour is essential in the communication of goals and intentions that leads to successful social cooperation.

There are two main problems here, both of them anticipated by McGeer‘s arguments for her agency theory of self-knowledge. One is that the traditional picture isn’t really well-equipped to explain how it is that we are so confident in our psychological ascriptions, beyond providing the obvious answer that these practices seem to be working fine so far. As shown, in McGeer’s alternative picture the reason folk-psychological practices work so well is because they are based on expressions of intentional states made as normative commitments. People are assumed to have special authority over their psychological states because they are constantly active in shaping these states. Note that as McGeer emphasizes this is not the same as saying that people cannot be mistaken when reporting their psychological states. There is other evidence for their reports that can be relevant to folk-psychological practices and there is, as mentioned, the possibility that people can be mistaken about themselves to the point where they are, in a way, taken out of commission. Seeing things this way, when taking part in folk-psychological practices people have the responsibility to be motivated by their normative commitments in order for others to trust them in a social context.

The limits of the traditional view of folk-psychological practices have also been stressed by Daniel D. Hutto and Tadeusz W. Zawidski, whose work can also provide some support for McGeer’s take on what these practices entail[10]. According to Hutto, theorists who focus on just the predictive and explanatory aspect of folk-psychology miss out on the fact that frequently there is no need to engage in theorizing or simulation in order to explain one another’s actions, since we can tell stories about our actions that can reveal the reasons behind them without often needing to engage in the kind of activities described in the common approaches to folk psychology. It is our capacity to exchange these justifying narratives that underlies our social interactions, Hutto argues, and not any specific theory or simulation-based affinity on our part for accurately predicting and explaining human behaviour. For Hutto,

[10] See Hutto, 2004, 2007, and Zawidski, 2008.

> "'folk psychology' is an instrument of culture, giving us the grounds for our evaluative expectations for what constitutes good reasons. This is not the same as merely providing a framework for disinterested explanation and prediction." (Hutto, 2004, p. 559)

His stance is compatible with McGeer's in that in his view as well, we don't need to look at how successful we are at accurately predicting and explaining one another's behaviour from an external standpoint in order to account for the pervasiveness of folk psychology in our lives. What makes folk psychology unique for us is that it enables us to justify each other's actions by reference to our reasons for acting and to hold each other responsible for offering such justifications and being able to live up to them[11]. To put this more in line with McGeer's stance, learning to engage in folk psychology involves more becoming skilled in making ourselves intelligible from a folk-psychological perspective that depends on the common understanding and expectations we share of intelligible behaviour, and less becoming skilled at telling what the mental factors issuing in our behaviour are. Here's what the skills involved in these practices might consist in, according to McGeer:

> "First, there are skills involved in saying and doing what is generally regarded as normal, reasonable or expectable in context- knowing how to negotiate the complex norms that govern so many aspects of our social-communicative lives. And here the narrative structures of sense-making folk-psychology have a role to play in establishing and reinforcing "canonical" patterns of behaviour: By way of these narratives, we learn what "reasonable" actors will think and do in a variety of situations. Still, reasonable actors are not limited to thinking and acting in canonical ways[.]…[T]here are skills related in being transgressive as well-specifically, skills relating to the asking and giving of reasons for untoward behaviour that still manage to place such behaviour within the sense-making ambit of folk psychology. Here the folk-psychological practice of attributing various psychological states finds a new role to play, not just in establishing what *is* canonical, but in negotiating what may count as reasonable

[11] See Hutto, 2004, p. 565: "[T]the traditional picture is only attractive if we assume that in giving explanations we always occupy an estranged, spectatorial point of view. Yet, in ordinary cases the other is not at arms length. For this reason the *standard* way we come to determine the reasons for which others act is dramatically unlike that employed in forensic investigations that seek to locate the cause of particular events. We cannot use the same sorts of methods we would deploy in determining, say, the cause of a plane crash. Rather, we usually rely on the revelations of others. They explain their actions for themselves. Of course, their admissions are defeasibile and often people are self-deceived about their reasons for acting. But we have fairly robust methods for testing, questioning and challenging such aims when it is important to do so, as in legal cases. We compare one person's avowals with the accounts of others, uncovering lies or internal contradictions that will invalidate either their testimony or their credibility. Countless everyday conversations involving the explanation of actions in terms of reasons mimic this process to a greater or lesser degree."

> even while departing from what is normally expected."[12] (McGeer, 2007b, p. 148)

Zawidski is another theorist who elaborates on this line of thought, by arguing that in order to explain how folk-psychological practices have persisted for so long we need to understand their function of setting a standard used as a guide for the ways in which we can intelligibly behave. This is what he calls the mind-shaping function of folk psychology, which he distinguishes from its mind- reading one. One of his main arguments for the prominence of this aspect of folk psychology is that if folk-psychological practices had the main function of identifying the mental states expressed in our behaviour and anticipating their effects, the fact that they have persisted for so long and have not been eradicated during our evolution as a species seems particularly puzzling, considering the limits in engaging in such practices. To illustrate this point, he makes an analogy to traffic rules. If all drivers had to actively predict and interpret each other's behaviour from moment to moment in order to cooperate, chaos would ensue before too long.

Fortunately, there is a framework wherein all drivers interact, which supports their attempts to anticipate and explain each other's actions. This framework is established through the common understanding of these rules that competent drivers display and expect others to share with them. The fact that all drivers are expected to conform to this shared understanding of their situation makes anticipation and explanation of driving actions easier for them, assuming they are motivated to do their best in respecting the basic rules of traffic that constitute their shared understanding[13]. As drivers conform to a basic understanding of traffic regulations, so do human agents in general conform to the norms inherent in their folk-psychological practices. In Zawidski's words:

> "[E]volution discovered simple mechanisms for shaping hominid behaviour so as to make it more predictable, or at least easier to coordinate with. Among these was the practice of ascribing propositional attitudes defined by normative

[12] For the idea that we use narratives to understand and justify each other's actions, see also Jerome Bruner, 1990, *Acts of Meaning,* Harvard University Press, whose work has influenced both McGeer and Hutto's theoretical standpoints.

[13] See Zawidski, 2008, p. 199: "There is no way that we can divine the cognitive states of fellow drivers, in the heat of traffic, with sufficient speed and accuracy to avoid catastrophe. Fortunately we do not need to. This intractable epistemic task is off-loaded onto our social environment. Legislatures pass laws and educators teach novices in such a way that the coordination problem becomes exponentially more tractable."

> relations to each other […], to which, thanks to various mechanisms of socializations, hominids strive to conform. Because of this, solutions to coordination problems do not depend on reliably accurate predictions based on correct ascriptions of cognitive states- an epistemically intractable task. Rather, they depend on figuring out what the *normatively sanctioned* response to some problem is, and assuming that others do the same […]. This assumption is justified by the efficacy of mechanism and practices of mind shaping." (Zawidski, 2008, p. 199)

The second main problem for the traditional picture, which is also brought to bear on the discussion by the proponents of our alternative account, is that adopting this view seems to make the practice of positively or negatively evaluating agents for their actions redundant[14]. When agents frequently fail to reliably conform to the predictions and explanations offered for their actions, for example, the traditional picture seems to imply that these predictions and explanations are somehow flawed. If failing to understand certain agents' actions was only a matter of displaying various inaccuracies in our folk-psychological attributions, for example, then it seems that the proper response to such failures would be to change our attributions so that they more accurately reflect the behaviour we are trying to account for. But then there would be no need to confront the objects of our folk-psychological understanding for failing to live up to this understanding, since it seems that the fault would be with our way of seeing things and not with their behaviour. Positive and negative evaluations seem out of place in such a context. We can further clarify this point by taking Zawidski's elaboration of his traffic analogy. As he points out, when a driver makes a mistake, the traffic regulations don't usually change to accommodate this mistake. Instead, the driver is confronted for failing to live up to the common expectations shared by all competent drivers.

Our alternative viewpoint is better-equipped to deal with this problem, since according to it the reason humans are evaluated for their actions is that they engage in these actions based on the normative commitments they make. And these

[14] See e.g. McGeer, 2007b, p. 148: "This is one of the most telling features that differentiates folk psychology as a regulative practice from what it would be like if it were a mere explanatory-predictive practice, appropriately construed as a proto-scientific theory of behaviour. For in the case of a proto-scientific theory, failure in explanation and prediction should lead to some revision in the theory itself or in the way the theory is applied; it does not lead to putting normative pressure on the "objects" of theoretical attention themselves to encourage them to become more amenable to folk-psychological explanation and prediction on future occasions.

normative commitments, as is made clear in McGeer's picture, go hand in hand with a variety of privileges and responsibilities. Competent agents know that they are viewed as such by others and also expect others to share the same viewpoint in all their social interactions. This leads to negative and positive evaluations of their own and others' actions and in corrective steps being taken for agents who frequently fail to live up to these common expectations and to make themselves understood by competent practitioners of folk psychology. Human beings can be criticized when their psychological claims don't cohere with their circumstances and their behaviour and steps to create a more coherent picture out of these elements can be taken. Agents might be able to take these steps by themselves and others might be able to help these agents take these steps by altering their circumstances in certain ways (by bringing these failures to their attention for example). If we have this kind of understanding in mind, which highlights the agent's role in living up to folk-psychological attributions, such a collective practice of positive and negative evaluations fits quite well with the practices of interpretation, prediction and explanation folk-psychologists engage in.

Paying closer attention to this fit between the normative and predictive-explanatory folk-psychological practices brings us to another issue that is worth examining in outlining our alternative conception of folk psychology. I am calling the view that folk psychology functions as a normative tool an alternative to the traditional picture to highlight our departure from a focus on just our capacity to predict and explain each other's actions by a combination of simulating and theorizing about our mental states. However, this should not be taken to imply that our alternative entails that we never exercise the latter capacity. Our interactions in a folk-psychological framework might well occasionally involve having to predict and explain each other's behaviour using a mixture of theorizing and simulating, as well as involving acting as reason-guided agents and expecting one another to act as such. Be that as it may, I take it that in the cases that we do engage in predicting and interpreting each other's behaviour by attributing intentional states to one another, it is the normative function of folk psychology that underlies any success we might have in doing so. It is because we learn to live up to each other's understanding of reason-guided behaviour by exhibiting certain behavioural patterns that manifest our intentional states that attributions of intentionality to one another work when they do and enable us to successfully anticipate each other's actions. This is the view that McGeer seems to hold[15], while Hutto and Zawidski also seem open to the idea that our folk-psychological practices are primarily regulative in their function but can also, for this reason, ground successful prediction and interpretation by the use of theorizing and simulation[16].

Finally, our departure from the traditional view of folk psychology can also be taken as a defence against the line that has been taken by some of the theorists looking at folk psychology (Paul Churchland's eliminative materialism is a well-known example[17]) that folk-psychological practices could be in principle replaced

[15] See McGeer, 2007b, p. 149:
"If we learn to govern our behaviour in ways that make us more readable to others, then their work as interpretative agents is greatly reduced. The same is true for us, if they learn to govern themselves likewise… We can, of course, show considerable interpretive ingenuity when called upon to do so; and this may require drawing upon fairly generalized knowledge about the psychological springs of human behaviour in addition to whatever particular knowledge we may have of individual peculiarities. However, what is exceptional of these moments in not only their relative infrequency, but also the difficulty and uncertainly with which such interpretive efforts proceed."
See also *ibid,* p. 150:
"When we develop as folk psychologists, we no doubt hone our interpretive skills; but, more importantly, we come to live in a world where the kind of interpretive work we need to do is enormously enhanced by how much meaning our interactions already carry for us and carry because of the way we habitually conform to norms that invest our actions with common meaning."

[16] Hutto, similarly to McGeer, argues that engaging in predictive-explanatory practices is still an aspect of folk-psychological expertise, even if it isn't at the core of our social interactions. As he makes clear,
"we may be forced to make predictions and explanations of actions precisely in the sorts of cases in which we do not know what to expect from others or we cannot engage them directly. But for this very reason these sorts of approach are bound to be, on the whole, much less reliable than our second-person modes of interaction." (Hutto, 2004, p. 565).
Zawidski also considers such a view, admitting that the extent to which it is accurate is an empirical question. See Zawidski, 2008, pp. 204:
"[I]t is possible that mind shaping functions to socialize individuals such that they are more likely to token the kinds of propositional attitudes, and engage in the kinds of behaviors that their typical interpreters expect. On this view, accurate descriptions of mental states supporting accurate predictions of behavior remain central functions of propositional attitude ascription. However, propositional attitude ascriptions succeed in realizing these functions only to the extent that they also succeed in prior shaping of individuals, to make 'abnormal' propositional attitudes and behaviors less likely in populations of interactants."
See also ibid, p. 205:
"Nothing in the foregoing is meant to suggest that human beings do not predict each other's behavior, nor even that they never use mental state ascription to this end. Once the use of mental state ascription to mind shape is reliable and prevalent, a derivative mind-reading use is possible, much as we predict that motorists will stop at red lights. The more effective mechanisms of socialization are at molding individuals capable of and willing to conform to the norms of folk psychology, the easier it is to predict individuals in such terms."

[17] See Paul Churchland, 1981, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78, pp. 67-90.

with a more accurate theory which will be better informed by scientific advances in areas such as neuroscience. As Mc Geer notes, such arguments do not take in consideration the normative and training role folk psychology plays in our everyday interactions. If the replacement of folk psychology with a more accurate scientific theory would mean that the agent's first-person perspective on his behaviour and talk of responsible agency wouldn't be part of this theory, then that might count as a reason to reject it:

> "For though its detractors might like to argue the point, the project of replacing folk psychology with a more "scientific" way of understanding ourselves is not in principle doomed to failure. But, pace the advocates of such reform, I do not think that such changes would be demanded by so-called "facts" about how minds work. Minds are as much made as discovered." (Mc Geer, 1996, p. 512)

Folk psychology as the playground of agency

The view of folk psychology we are led to, for which McGeer's views serve as the main template, is one in which the agent's own active role in his behaviour plays a central role in the practices of interpretation, prediction, explanation and evaluation of intentional behaviour. The agent's active role in his behaviour is based on his self-understanding and on the normative commitments he makes, relative to this understanding. Folk-psychological activities are essential not only for the stable interaction between agents in social contexts, but also for the development of the very agency exercised in these interactions. Folk psychology in this sense is also used in teaching to developing agents how it is that they can best express their active role in their behaviour and the implications their exercises of agency have for themselves and others. In other words, the assumption is that by learning to think in a folk psychological context from an early age, humans learn to explore the implications of having certain beliefs, desires and other intentional states and they also learn to hold themselves and others responsible for the expression of these intentional states in their behaviour. They also learn to express these intentional states in a coherent manner, in order to be better understood by other agents who attempt to anticipate and explain their behaviour. And in turn they expect the same coherent expression of agency in behaviour from everyone else they are cooperating with in a social context.

In fact, folk-psychological training can be discerned in our everyday lives quite easily. One relevant example comes from McGeer and Philip Pettit's "The Self-Regulating Mind":

> "Consider the well-known phenomenon of "parental scaffolding"[18] or the over-interpretation of children's "intentional behaviour"…. Little Susie happens to put her toy rabbit in bed and we tell her that that's right, she should look after the rabbit and keep it warm, because it's her friend. Or she unthinkingly gives her baby brother a sloppy kiss on the forehead and we tell her what a good girl she is to love him like that and to want to show him that she loves him. Or she parrots a teacher's claim that New York is a big city and, perhaps correcting other claims she makes about New York, we tell her how clever she is to know that and try to get her to understand the claim by treating her as if she had made it with full understanding: for example, by adding that she is pretty clever to know that New York has so many more people living there than in her home town." (McGeer and Pettit, 2002, p. 297)

Receiving similar training, we learn for example not to believe contradictory things, to believe the consequences of our beliefs, to behave according to a certain combination of beliefs, desires and other psychological characteristics expressing ourselves, and to recognize that acting in such a way as competent agents (acting for reasons) brings with it certain privileges and responsibilities by which we must abide if we want to be taken seriously in a common folk-psychological context. In some sense then, engaging in folk psychology can be described as a game in which we all learn to take part in from a very early age and which supports the various roles we subsequently assume within a social/cooperative framework. Assuming these roles entails that we all learn to treat each other's behaviour as comprising the actions of reason-guided agents and to live up to this image.

In the remainder of this thesis, I intend to demonstrate that the alternative conception of folk-psychology that I've delineated in this chapter can help with accounting for different aspects of human agency and with solving a variety of puzzles, of the kind identified in our introduction, that are associated with developing such an account. In the next chapters, I will be mainly looking at what it means to act as an agent in a human society and at the way in which the nature of our agency is also shaped by the normative constraints inherent in the common understanding of agency that we share with other agents. As we'll see, we can make significant

[18] This term originates in Jerome Bruner, 1983, *Child's talk: learning to use language*, Norton, New York.

headway in explaining the nature of our capacity to express ourselves authoritatively in our actions in a self-knowing and self-controlled manner if we place this capacity in the context of our social interactions, which depend on a constant exchange of reasons in support of our actions. My main objective is to develop a promising account of human agency within a folk-psychological setting by mainly focusing on perspectives from the philosophy of action and mind, while still respecting more empirically oriented viewpoints from areas such as cognitive science and neuroscience.

Chapter 2 mainly deals with the nature of self-knowledge and with our capacity to express this knowledge in our actions. I argue that our self-knowledge is constituted by the normative judgments we make and that we use these judgments to regulate our behaviour in accordance to our folk-psychological understanding of agency. We are motivated to act as such because of our motive to understand ourselves, which has developed through our training as self-knowing agents in a folk-psychological framework.

Chapter 3 explores the idea that we develop a self-concept which enables us to act in a self-regulating manner. I distinguish self-organization from self-regulation and argue that we are self-regulating in our exercises of agency because we have developed a self-concept that we can express in our actions. What makes us distinct from other self-regulating systems, however, is that we can also recognize and respond to the fact that being such systems brings us under certain normative constraints and that we have to interact with others who are similarly constrained.

Chapter 4 is mainly concerned with placing empirical evidence which illustrate the limits of our conscious awareness and control in the context of our account of agency as a complex, emergent social phenomenon. I argue that even though it is empirically plausible to suggest that our actions are mainly shaped by processes we are not consciously aware of and do not consciously control, this should not be taken to imply that we are unable to act as agents. That is because our engagement in the practice of a constant exchange of reasons with other agents is more crucial for our agency than being able to consciously initiate and control our actions.

Finally, chapter 5 deals with the way in which agentive breakdowns such as self-deceptive inauthenticity fit with our account. I mainly focus on cases where we are somehow self-deceived about what we're doing and argue that they should be understood as cases in which we fail to properly express ourselves in our actions by either failing to offer reasons for our actions or offering reasons that are unsatisfactory. In keeping with the theme of my thesis, my view is that the extent to which this is the case and what the proper response to these kinds of failures should be also depends on the folk-psychological framework wherein our collaborative exercises of agency take place.

Chapter 2

The Search for a Common Thread: Self-Knowledge and Agency

Introduction

I take it that in the course of interacting with their environment and each other, human beings express their agency through achieving a certain kind of self-understanding and expressing this understanding in their behaviour. Crucially, this kind of self-understanding depends on the development of a self-concept consisting of various attributions of intentional states such as beliefs, desires, intentions, hopes, fears, traits, attitudes, habits and other similar mental characteristics. Presumably, a combination of such mental states consists in the acting individual's motives for acting in a certain way. A different way to view the self-concept developed by rational agents then is as a reflection of the mental motivating factors that lead them to act in certain ways.

This self-concept can be genuine and reflect the acting individual's actual mental states or it might be a product of faulty understanding of one's mental composition and consequently refer to intentional states that do not correspond to the main motives the individual has for acting in a certain way. In this chapter, I will not dwell on what exactly it means for self-understanding to be genuine or to be faulty in some way and on the consequences a faulty self-understanding has on what the correct treatment of the agent displaying such understanding should be. What I will focus on is the development of an account in which human beings display their rational agency through actively expressing their intentional states in their actions, and on what such an expression would consist of in ideal circumstances in which rational agents express themselves in their behaviour successfully.

In developing this account I will mainly draw from three sources. The first is Victoria McGeer's theory of self-knowledge viewed as depending on the agent's adopting certain normative commitments influencing the ways he behaves[19]. The

[19] See Victoria McGeer, 1996, "Is Self-Knowledge an Empirical Problem? Renegotiating the Space of Philosophical Explanation", *Journal of Philosophy* 93, pp. 483-515 and McGeer, 2007a, "The Moral Development of First-Person Authority, *European Journal of Philosophy* 16(1), pp.81-108.

second is J. David Velleman's development of a view in which the desire for self-understanding is the underlying motive constituting agency and human beings express themselves in their behaviour through incorporating their self-concept, their actions and the circumstances in which they find themselves in a coherent narrative[20]. The third is Richard Moran's exploration of what he calls the "deliberative" stance, which depends on the agent's first-person perspective on himself and his behaviour and is distinct from what he calls the agent's "theoretical" or "empirical" stance, which consists of a third-person, external viewpoint on one's mental composition and one's behaviour[21].

Through a combination of elements from these viewpoints I intend to develop an account in which the agent's first-person perspective plays a crucial role on his understanding of himself and on his various expressions of agency. Such an agent expresses his active contribution in his behaviour and performs certain actions by developing a self-concept consisting of his self-attributions of intentional states and expressing this self-understanding in his behaviour. In this framework, the self-attributions the agent makes function as normative commitments for him, prompting him to shape his behaviour in such a way that it ends up reflecting the way he sees himself. And the way the agent sees himself is based on these self-attributions, which can also be influenced by external attributions of mental states made to the agent by observers of his behaviour. So the agent's self-concept is based on his self-attributions, which reflect the agent's unique first-person perspective and the agent's normative commitments to act in certain ways that make this first-person perspective more salient both to himself and to the persons he interacts with in a social context.

Furthermore, the agent's desire to understand himself and his actions as resulting from his own viewpoint on the world and to convey this understanding to the other agents he comes in contact with is precisely what provides the normative force to the

[20] See J. David Velleman 1992, "What Happens When Someone Acts?", *Mind 101,* pp.461-481, Velleman, 2009, *How We Get Along,* Cambridge University Press and Velleman, "The Self as Narrator", Available at http://www.uwm.edu/~hinchman/Velleman-Dennett.pdf

[21] See Richard Moran, 1997, "Self-Knowledge, Discovery, Resolution, and Undoing", *European Journal of Philosophy* 5 (2), pp. 141-161, Moran, 1999-2000, "The Authority of Self-Consciousness", *Philosophical Topics,* pp. 179-200 and Moran, 2001, *Authority and Estrangement: An Essay on Self-Knowledge,* Princeton, NJ: Princeton University Press.

agent's self-attributions constituting his self-concept. Finally, the external and self-attributions made by these interacting agents depend on their sharing a common folk-psychological understanding of what it means to be an agent and of the ways such agents interact with each other and with their environment. In unpacking this account I will start with providing a more detailed explanation of what it means for agents to adopt certain normative commitments to act in various ways expressing themselves.

The binding force of self-attributions

The idea that an agent is shaping his own behaviour because he is expressing his own intentional states in it implies that he is somehow aware of these states in the first place. If one is aware of his own psychological constitution (or of parts thereof) then it can be said that one has a certain measure of self-knowledge. Self-knowledge, of course, is yet another tricky concept whose meaning needs to be further clarified. The question regarding self-knowledge that is central to our purposes of providing an account of agency is this: What constitutes an agent's authoritative self-knowledge of his own intentional states? A highly promising answer to this question has been offered by Victoria McGeer. The main details of her view of what an agent is said to be doing when he displays self-knowledge of his intentional states have been sketched out in the previous chapter.

The bottom line of her view is this: an agent is said to be aware of his own intentional states because he plays an active role in displaying certain patterns of behaviour which express these states. These patterns of behaviour can range from specific emotive responses and thoughts to more public displays of behaviour which consist in this individual's actions. McGeer understands mental states such as beliefs and desires as dispositional, in that the agent is disposed to feel, speak, think and generally behave in ways expressing his hopes, fears, intentions, beliefs and other such intentional states. Since he is the one actively expressing these states in his behaviour, he is intimately aware of these states in a way that someone observing him is not. That's because the agent can attribute certain thoughts, feelings and actions to himself and he can fit these self-attributions to his behaviour in order to validate them. McGeer argues that we should understand the agent's self-attributions as an agent's normative commitments because the agent is motivated to act in ways that can be understood in the context of these commitments[22]. They can more properly be understood as judgments that the agent makes on what patterns of behaviour are appropriate in his given circumstances, which he can also express publicly.[23] These normative commitments enable the agent to speak authoritatively for what his psychological states are, and so to be intimately aware of them in the manner of self-knower, because he is the only one in a position to display the dispositions expressed in them[24]. McGeer describes such intimate knowledge as the knowledge a driver of a car would have as opposed to its passenger[25]. Sure enough, the passenger can observe the car's movements and come to certain conclusions about it (for example, come to know where this car is heading and with what speed it is heading towards its destination) but the driver is aware of these facts because she is the one who makes them true. She drives the car towards a certain destination and with a certain speed, and thus has first-person knowledge of these facts.

In this same way, in McGeer's point of view, an agent who is compelled by certain normative commitments that he's made is aware of his own mental states expressed by these commitments. Others can also come to know the agent's mental states through observing his patterns of behaviour, which also depend on the agent's efforts to live up to his normative commitments. The agent's normative commitments are essential to the behaviour that others can observe, since the specific behavioural patterns on display would not be the same had the agent not had the normative commitments that he is motivated to live up to. This fact can help explain why the agent's claims of self-knowledge have a different status from other's claims of being aware of the agent's psychological constitution. The agent is guided by his

[22] The source of the normative force of the self-attributions the agents make has been hinted at in Chapter 1 and will be explained in more detail in the following discussion. See, for example, this chapter's discussion on our drive for self-understanding and its role within a folk-psychological framework.

[23] Eric Schwitzgebel has argued for a similar distinction in the case of belief, between commissive judgments and the dispositions expressed by those judgments. See Eric Schwitzgebel, 2001, "In-Between Believing, *Philosophical Quarterly* 51, pp.76-82, Schwitzgebel*,* 2002, "A Phenomenal, Dispositional Account of Belief", *Nous* 36, pp. 249-275 and Schwitzgebel, 2005, "Acting Contrary to our Professed Beliefs", available at http://www.faculty.ucr.edu/~eschwitz/

[24] See McGeer, 2007, p. 82: "The agent has a privileged authority in self-ascribing intentional states because it is she who makes it the case that she deserves to be ascribed these states; she has "maker's knowledge", not the knowledge of a particularly accurate perceiver or detector".

[25] See McGeer, 1996, p.505: "The privilege of first-person knowledge is thus really more like the knowledge of a person driving a car as opposed to her passenger. The passenger may very well see where the driver is going, but still does not know in the immediate *executive* sense of the driver herself".

understanding of himself by making an effort to live up to his understanding, whereas other observers' interpretations of this agent's behaviour do not typically influence his behaviour in the same way. As McGeer is careful to note, this does not imply that the agent is infallible about his own mind. In certain cases an observer might be in a position to know more about the agent's mind than the agent himself (if, for example, the agent is systematically falsifying the normative commitments he undertakes by behaving in completely unrelated ways from those that he is claiming [in public and in private] that he will act)[26]. But, as McGeer argues, these cases are not the norm and we, in our social interactions, are inclined to give each other a significant leeway to adjust both our normative commitments and the rest of our behaviour.

On this understanding of agency, the agent's privileges and responsibilities are up to him in a more fundamental way than if they were simply stipulated as a matter of social convention. I think this is an advantage of McGeer's view, since it provides a more solid foundation for our practices of evaluating and correcting each other's intentional actions. Since the agent has to live up to his status as being in control of his own behaviour by matching his normative self-ascriptions with his behaviour, it seems to me to make more sense to criticize him when he demonstrates particular discrepancies between these two elements and when he regularly says one thing about himself and does something completely different. I also think it makes more sense in this case to say that the agent can achieve a better self-understanding due to certain corrective practices, because it is assumed that it is up to him to learn to produce a better fit between his behaviour and his normative commitments[27].

The unreliability of introspection

McGeer's answer to the question of what it is that constitutes the agent's authoritative self-knowledge of his intentional states seems to be at odds with a different approach to explaining the agent's grasp of his mental states. This approach might claim that the answer to the question of what constitutes an agent's self-

[26] See Chapter 5 for a discussion of when an agent's claims of self-knowledge might be false.

[27] For a similar account of the role of agency in understanding self-knowledge, see Akeel Bilgrami, 1998, "Self-knowledge and Resentment", in C.Wright, B.C.Smith and C.Macdonald (eds.) *Knowing Our Own Minds,* Oxford: Oxford University Press. Bilgrami shares with McGeer the conviction that our common practices of holding each other responsible for acting as self-knowing agents by living up to our claims of self-knowledge are essential for our capacity to be intimately aware of our own intentional states.

knowledge of his intentional states lies in the fact that the agent stands in a privileged epistemic standpoint in respect to these states because he can accurately recognize and report them. His self-knowledge in this case might be based on a special insight he has on the contents of his mind, an insight that external observers of his behaviour do not share with him because they do not share with him this special epistemic relation to his states (though they might still be able to divine these states using different means). This kind of inner knowledge that the agent possesses might be acquired through a variety of means, involving perception, inference and introspection. Perhaps he is able to accurately detect his intentional states by being in a good position to detect their effects on his behaviour and infer their influence in his actions, or by recognizing these intentional states through introspection, or perhaps even by having some kind of immediate awareness of their occurrence that doesn't depend on using any indirect means to garner their existence. An approach to self-knowledge that relies on one such special first-person relation to mental states as a starting point in order to provide an answer to the question of what constitutes the agent's authoritative self-knowledge of his intentional states, whether this relation is construed as depending on perception, inference, introspection or on immediate awareness, might have a lot to recommend it.

However, I will argue that such an approach is in fact misleading and falls short of providing a full answer to our question. This is in part because such an approach, at least when it argues for self-knowledge arrived at through observation, inference or introspection, might fail to take under consideration the ways in which our inner awareness of our mental states is fallible and inaccurate. Furthermore, even if such a construal of self-knowledge manages to avoid this problem, the deeper and more important reason against using it as a starting point for an account of agency is that it undermines the role of the agent in action and it fails to adequately account for the way in which authoritative self-knowledge is linked to agency. However much this understanding of our relation to our content of our minds has to say about our inner awareness of mental phenomena such as sensory or proprioceptive states, it is not enough to explain the authoritative aspect of the agent's self-knowledge of his own intentional states.

So why does an approach relying on means such as introspection to account for the agent's self-knowledge underestimate the unreliability of these means? For the purposes of providing more concrete examples for why this might be the case, we can focus on certain influential studies conducted in the fields of cognitive and social psychology, which hint at the extent to which we are mistaken about what the causes of our behaviour are and what certain feelings we are aware of signify about our own states of mind. One often cited classic study of this sort is Richard E. Nisbett and Timothy DeCamp Wilson's "Telling More Than We Can Know: Verbal Reports on Mental Processes"[28]. Another is Daniel M. Wegner and Thalia Wheatley's "Apparent Mental Causation: Sources of the Experience of Will"[29]. It's worth examining some of the findings mentioned in these articles in a little more detail in order to illustrate why it is puzzling to simply assume that the agent can regularly come to know the contents of his mind by relying on means such as perception, inference or introspection.

Nisbett and Wilson conduct and review a number of different experiments in which subjects were asked to provide some explanations for what kind of stimuli influenced their behaviour, or what kinds of mental processes led to their behaviour. In most of these experimental settings, the majority of the subjects' proposed explanations of the causes of their behaviour did not correspond to what the actual causes were demonstrated to be. For example, one experiment conducted by Nisbett and Wilson had subjects watching an interview with a college teacher and reporting how attractive they found this person's appearance, mannerisms and accent. The group of subjects was divided in two, with half of them seeing the teacher answer questions in a friendly and approachable manner and the other half seeing him answer in a reserved and cold manner. As expected, the latter group of subjects found the teacher's appearance, mannerisms and accent less attractive than the former group. When asked if what they thought of the teacher in general affected the way they rated his physical attributes, all subjects denied that and provided the exact opposite causal

[28] See Richard E. Nisbett and Timothy DeCamp Wilson, 1977, "Telling More Than We Can Know: Verbal Reports on Mental Processes", *Psychological Review* 84 (3), pp. 231-259.

[29] See Daniel M. Wegner and Thalia P. Wheatley, 1999, "Apparent Mental Causation: Sources of the Experience of Will", *American Psychologist* 54, pp.480-492.

explanation. They claimed that their dislike (or like) of his physical attributes influenced their dislike (or like) of his character.

On another experiment reported by these authors and conducted by Nisbett and Schachter[30], subjects had to take a series of electric shocks, with some of them being given a pill said to produce symptoms that the experimenters knew were similar to the symptoms produced by electric shock (e.g. breathing irregularities). In fact, the pill had no such effect but the subjects who were given this placebo were able to withstand more intense electric shocks than the ones who hadn't been given the pill. The experimenters' assumption was that the subjects taking the pill felt the symptoms induced by the shocks but attributed them to the pills instead, with the effect that they were able to withstand these symptoms a lot longer than other subjects. The interesting part here is that when the subjects who took the pill were asked why they were able to withstand the shocks to the extent they did, only three out of twelve subjects mentioned thinking of the effects as produced by the pill and not the shocks. The rest of the subjects, as the experimenters assumed, had no idea that this is what they were doing and instead offered all kinds of different causal explanations for why it was that they behaved in the ways they did.

In the case of Wegner and Wheatley, their main concern is with the feelings people have when they believe to be consciously willing their actions. These authors performed a study which indicated that people can be led to erroneously believe that they are in control of their behaviour and that they have a specific intention to behave in a certain way, when in fact their behaviour had been externally manipulated. The primary study they used to demonstrate this hypothesis was what they called the "I Spy" study. In this study, subjects were asked to monitor a computer screen containing various objects and move a mouse pointer to whichever object they chose when a specific cue was given to them. These subjects were cooperating in moving the cursor with another person who, unbeknownst to them, was instructed to force some stops near a specific object. Before these stops the subjects heard an object being named through their headphones, which usually ended up being the object closest to the pointer when the stops had been forced. Also, the subjects were allowed to make stops on their own as well, in which it was established that they

[30] The study cited is Richard E. Nisbett and Stanley Schachter, 1966, "Cognitive Manipulation of Pain", *Journal of Experimental Social Psychology* 2, pp. 227-236.

didn't tend to stop the mouse cursor closest to the object they heard named through their headphones. This indicates that they were unlikely to have actually intended to stop the cursor where the insider did when the stops were being forced.

What actually transpired though is that when the subjects were cued with a specific word shortly before the insider forced a stop, the subjects reported on average that they had intended to stop the cursor near that object. The shorter the time span before the auditory cue received by the subjects, the more these subjects felt that they had intended to make a stop. Wegner and Wheatly use these results in conjunction with other considerations to argue that one's feeling of consciously willing an action is not an accurate indication of the underlying process taking place. They go on to claim that conscious will is an illusion since unconscious mechanisms are responsible both for the thoughts accompanying an action and for the initiation of the action itself. I believe that this conclusion is overly ambitious and that the results examined by Wegner and Wheatley do not suffice to establish such a strong claim, but I will not argue for this here[31].

The point that I want to make here and which is also argued for by these authors is this: introspection is overrated[32]. What these authors notice and is of particular interest for our purposes is that the subjects of such experiments cannot be plausibly said to engage in a process of introspection when they offer causal explanations of their actions and when they report on what they actually intended in those circumstances. This doesn't change the fact that they still were pretty comfortable in talking about what was going on in their own mind and acting as if they had some kind of privileged first-person authority, as if they actually knew what they believed for example and what they were doing in a given circumstance. This would be puzzling in these cases if we simply relied on them having a reliable introspective access to the contents of their own mind. But if we adopt a view in which people know what they're thinking and what they're doing because they try to present themselves, publicly and privately, as authoritative and competent rational agents,

[31] See Chapter 4 for a more extensive discussion of the interplay between conscious and unconscious processes and its role in our expressions of agency.

[32] See also Eric Schwitzgebel, 2008, "The Unreliability of Naïve Introspection", *Philosophical Review* 117, pp. 245-273, in which he argues that our introspection might be unreliable even in cases where it seems it couldn't go wrong, e.g. when introspecting on our current thoughts. His basic claim is that "we're prone to gross error, even in favorable circumstances of extended reflection, about our ongoing emotional, visual and cognitive phenomenology." (Schwitzgebel, 2008, p. 259).

then this readiness to talk about what goes on in their mind becomes less puzzling. Once again, this doesn't mean they're always right. But it does mean that they are in a position to show that they have an authoritative knowledge of their intentional states, because they can act in accordance to their claims of self-knowledge, even if they do not possess a reliable introspective access to the contents of their mind.

In the case of these particular experiments, this explanation may seem inadequate. Nisbett and Wilson offer a different explanation for why it is the subjects talked about the causes of their behaviour and their mental lives with such confidence. They claim that people in such circumstances make some assumption about what the most plausible cause of their behaviour was, based on the most influential implicit causal theories that they possess. As Nisbett and Wilson argue, there might be some characteristics of certain stimuli received by the subjects which make them particularly good candidates for these subjects as causes of their behaviour[33]. For example, if a stimulus is particularly salient to a subject and there are no other possible causes of his behaviour that he is aware of, then this subject is likely to postulate that stimulus as playing a role in causing his behaviour. And if this subject believes that there is a high correlation between stimuli of this sort and certain kinds of behaviour, then his confidence in seeing it as a relevant cause will probably be higher.

I don't deny that this is part of what's going on when subjects make mistaken attributions to themselves under the aforementioned circumstances. This kind of explanation sounds like what a theory-theorist about folk psychology would say that people do in general when explaining and predicting each other's behaviour. But as we have seen in the previous chapter, a theory-theory perspective on its own does not suffice to explain all the functions of folk psychology. People might use an implicit theory about how a human mind functions not only to explain, interpret and predict each other's behaviour but also to develop their own particular world-view and the way they see themselves. The perspective they develop though exercises an important influence on the actions they perform as agents and it is not just an abstract

[33] See e.g. Nisbett and Wilson, 1977, p.255: "Confidence should be high when the causal candidates are (a) few in number, (b) perceptually or memorially salient, (c) highly plausible causes of the given outcome (especially where the basis of plausibility is an explicit cultural rule), and (d) where the causes have been observed to be associated with the outcome in the past".

theory which does not necessarily correlate to anything in the way their minds actually function. These experiments then might indicate the subjects' use of assumptions of plausibility and their use of certain causal theories when asked about their actions in these particular experimental scenarios, but they do not suffice to explain what it means to act as an agent in general. However, I think that they are useful as hints towards the limits of introspective access to one's mind.

The view from inside

As alluded to earlier, there is a deeper and more conclusive reason to reject an approach to the agent's self-knowledge of his intentional states that primarily focuses on a privileged epistemic access the agent has to these states. The reason is that this kind of approach, even if it construes the privileged epistemic access the agent has to the occurrence of his own intentional states as an access that is immediate and especially reliable, so even if it manages to avoid the problems related to introspective access to the contents of one's mind, for example, is still problematic as an answer to what constitutes the agent's authoritative grasp of his intentional states. That is because this kind of approach displays a serious misunderstanding of the role of the agent's first-person perspective on his own intentional states and its importance for providing a link between the agent's self-knowledge and his agency. The importance of this role has been hinted at with McGeer's car example, in which the driver exercises a different kind of control on the car's movements and has a different kind of knowledge of them from its passenger. The agent, similarly, exercises a different kind of control on his attitudes and has a different kind of knowledge of them from an external observer of his behaviour. He has a view from the inside, as it were, not because he can peek at the intentional states occurring in his mind but because he acts in ways that express these intentional states. Here is what McGeer has to say about the subject:

> "….in presenting us as creatures simply assailed by a conscious awareness of our first-order states, we are unwittingly presented as utterly passive, not in control of our various thoughts and action, and so not able to take responsibility for them. To be viewed properly as agents, we must be construed instead as actively involved in forming, reviewing, revising, suppressing, and selectively acting on the first-order states we "know" about because we are the ones generating those very cognitive processes."(McGeer, 1996, p. 505)
>
> "….in making claims about one's own cognitive and emotional situation, one is making claims about a situation, both internal and external, that one has

> played (and continues to play) an active role in creating and maintaining. Hence, this………would make sense of the doctrine of first-person authority, even in the face of occasional (perhaps, even, systematic) error in particular first-person utterances." (ibid, pp.505-506)

The agent's first-person perspective on himself and his actions then is seen as being essential to his expressions of agency, but more has to be said on exactly what kind of control over and knowledge of his mind and actions this perspective affords the agent. Richard Moran has recently provided a number of influential considerations in favour of a view which brings the agent's own first-person perspective into focus, while explaining what this perspective consists of exactly[34]. One central tenet of his general view is that the agent's perspective on his own intentional states is fundamentally different from an external perspective on these states and it expresses a different kind of control over them. Moran frequently calls the agent's first-person stance the "deliberative" or "transcendental" stance and the external perspective from which the agent's states are viewed as empirical objects of study the "theoretical" or "empirical" perspective.

The perspective of the deliberator, in Moran's framework, is not a perspective from which the agent is clearly distinguished from his intentional states, which he treats as objects of observation occurring within his mind. Instead, the agent's deliberation constitutes the intentional states that express him as a person. By focusing on certain facts which count as reasons for being convinced of the validity of a certain proposition or on facts that count as reasons for pursuing a specific action, the agent is involved in shaping mental states such as beliefs and intentions, respectively. When the agent makes a judgment based on certain reasons he perceives, the agent has formed an intentional state. As long as the agent is convinced by these reasons, the agent maintains whichever intentional state is supported by them.

Let's take an example of an agent's belief that the sky is purple. In this case the agent simply perceives the sky as such and thus has a belief that it is purple. Assuming his belief is rational, Moran would say, he does not perceive the sky as being blue but detects the presence of a contradictory belief in his mind, which he's subsequently forced to accept as his own. Nor does he observe his behaviour as

[34] See Moran, 1997, 199-2000 and 2001.

indicating that the sky is purple even though his best judgment is that it's not, so that he reluctantly admits to a belief that the sky is purple. The agent knows his own mind effortlessly and is resolute in his knowledge, not so much because of any kind of introspection or reliance on mediating factors, but because he is committed to a particular viewpoint.

There are three clarifications that Moran is very insistent on making in order to further explicate his view and to avoid any kind of confusion on what kind of viewpoint he is committed to. The first is that the agent's own intentional states are not always the product of an explicit and reflective process of deliberation on certain reasons which support them. What is of importance is that were the agent to be asked about these reasons, he would be able to provide them in support of his acting in the ways that he does and having the beliefs, desires, intentions and other mental states that he does.

Here Moran, in a similar way to McGeer, is embracing a viewpoint in which an acting individual is not only treated as an authoritative, self-knowing agent for the purposes of the stable function of social practices, but is also expected to live up to this status. As he characteristically writes, "the special first-person accessibility of mental states seems something we not just *grant* to people, but something that is a normal rational expectation we make of them" (Moran, 1999-2000, p. 185). For him, competent rational agency is seen as "demand, rather than concession" (ibid), in that we expect each other to act as if we are in control of our own actions and of our own minds and we don't just interpret each other as such, indefinitely modifying these interpretations so that they fit our ideal of rational agency. Central to this view is the idea that the agent is expected to be in a position to justify appropriately his own actions and states of mind by invoking certain relevant reasons for them, and that the agent might be criticized when he is frequently unable to provide such competent justifications. I think that this conclusion of Moran's view fits in with the view we have been developing of agents adopting certain normative commitments and being expected (by themselves and their peers) to act in ways that justify these commitments.

A second important clarification Moran insists on concerning his view is that there is nothing more the rational agent needs to do, when he concludes in favour of certain considerations, in order to adopt an intentional state. If the agent would have to adopt an external standpoint to his mental states and treat them as empirical objects which he must somehow manipulate in order to produce them in him, then he would not, in this view, be expressing his active agency with respect to these states. For Moran, this kind of control is not constitutive of agency and it can lead to viewing an agent who regards his states from such an external standpoint as alienated from them, precisely because he does not direct his gaze outwards to the facts that are relevant to the support of such states but has to resort to manipulating them as he would anything else in the external world. An agent can induce a variety of effects on himself by using external means (for example, by arranging his circumstances in certain ways so that he'll come to adopt specific beliefs) but Moran denies that states arrived at by such means are products of the exercise of responsible, authoritative agency.

The last clarification Moran makes when it comes to his view, and one that we have dwelt on to some extent already, is that the first-person authoritative knowledge a rational agent has of his own mind is not based on an any epistemic access he has on the contents of his own mind, even if this access is extremely reliable and immediate. Self-knowledge would not be authoritative, as Moran makes clear, if it was based on the agent coming to know about his own mind without being able to provide any reasons for whatever intentional states he maintains, by treating them instead as occurring independently of his endorsement of any reasons for them. This would be so even if the agent came to know about his own mind effortlessly and instantly by being able to somehow immediately detect his beliefs and desires while they were occurring in him. Moran uses an example involving a psychotherapist who, instead of applying his expertise on others to identify their mental states, applies it to himself and learns to do so in an entirely efficient and effortless manner[35]. In his view, if this person was only attributing these states to himself without being able to endorse any reasons for having them in the first place, he'd still be alienated from them and he could not be said to know his own mind as its active author.

[35] See Moran, 1999-2000, p.189: "For we could just as well talk about the analyst himself as the one with the unconscious attitude of resentment, but now both his theoretical expertise and his skill at applying it enable him to attribute this attitude to himself more or less immediately, without any laborious theoretical inference from the behavioural evidence".

Another way to put this is that the agent does not know his own mind authoritatively as long as he's unable to identify with the reasons in support of the mental states constituting his individual psychology and to endorse these mental states as his own. So, for Moran, the kind of immediacy required for the kind of authoritative self-understanding that goes hand in hand with the agent being in control of his own mind and actions is not just any kind of immediacy, but it is immediacy based on a normative commitment the agent makes on viewing the world in a certain way. The more the agent has to resort to external means and adopt the theoretical perspective to maintain his intentional states, the more he loses touch with the objects of those states and the less confident he and others become in attributing these states to him. Another example coming from Moran (which he borrows from Sartre)[36] is the one of the gambler, who has a resolution to stay away from gambling but also has knowledge of his failures to do that in the past. The more the gambler sees his resolution as an impotent mental state which has proved inadequate in the past, the more he loses touch with the real reasons he has for staying away from gambling and subsequently, the more he further weakens his resolution. The gambler's failure, in Moran's viewpoint, is that he fails to realize that it should be up to him as an agent how strongly he feels against gambling and how determined he is to stay away from it, instead of viewing this state as something that will run its course regardless of any kind of contribution he makes.

Self-regulation and agency

As Moran's position has been developed so far, it is importantly similar to McGeer's view on the way normative commitments reflecting an agent's self-concept guide him to shape his behaviour as he deems appropriate based on his circumstances. McGeer herself wrote about the similarities and differences between her view and Moran's[37]. Let's take a closer look at what contrasting these accounts can offer us for our own account of agency.

First of all, both authors insist that the agent's own first-person perspective on himself and his actions is crucial for explaining what kind of control he can exercise over his own mental states and how he can express that control in his behaviour.

[36] See Moran, 1997, pp. 148-150.

[37] See Victoria McGeer, 2007a, "The Moral Development of First-Person Authority", *European Journal of Philosophy,* Volume 16(1), pp. 81-108.

They also both point out that a view which treats the agent's knowledge of his own mind and his control over it and over his behaviour as simply dependent on perceptual knowledge and on certain inferences he makes (or, as Moran would put it, an agent's only adopting the empirical perspective) faces significant difficulties. In its extreme version, such a view would lead to the agent not being able to identify with his own mental states and to make up his own mind through deliberation.

One of the crucial aspects that would be missing from such a view, according to both Moran and McGeer, is the agent's ability to go beyond any kind of inferences or evidence or perceptual knowledge when displaying his self-understanding in his behaviour, an ability for example to normatively commit himself to a given course of action or to the truth of a proposition. In Moran's own words:

> "At some point I must cease attempting to infer from some occurrence to my belief; and instead *stake* myself, and relate to my mental life not as of symptomatic value, but as my current commitment to how things are out there." (Moran, 1999-2000, 196-197)

The agent's active role as the author of his own mind and actions is what these authors insist on not leaving out from any account of self-understanding and self-control, and because of this the agent's own first-person perspective cannot be disregarded.

Another similarity between Moran and McGeer's views is that they explicitly discourage anyone from taking their arguments as advocating that the agent can make up any kind of mind he likes, depending on his whims. The proper way, according to them, to view how the agent makes those normative self-attributions which constitute the self-understanding that he expresses in his actions is that the agent responds to his environment in the most fitting manner, depending on the way he views himself and his circumstances:

> "It is not that we are free to pick and choose whatever psychological states suit us best. It is rather that we engage our reason to determine what is appropriate to think, desire, and feel given how we find the world and our situation in it[.]" (McGeer, 2007a, 88)

In this sense, when the agent displays genuine self-understanding in his behaviour, he responds to his environment in a manner that best exemplifies his self-understanding. How exactly the agent does that and what it means for him to respond in the most appropriate manner to his environment will be further elaborated as our account of self-understanding and agency develops. For the moment, I'd like to focus on one important disagreement between Moran and McGeer's views on what the exercise of agency in one's behaviour consists in.

This disagreement has to do with Moran's point that an agent does not have to further manipulate his mental states in any kind of way once he has concluded his deliberation about the facts in a given circumstance. As we have seen, for Moran maintaining one's mental states by using means other than focusing on one's reasons for having these states does not consist in expressing one's responsible, competent agency in one's behaviour. McGeer disagrees with this position and argues that there are plenty of cases where we need to maintain our intentional states by implementing a variety of external means, and in some cases this can be viewed as an expression of our agency in our behaviour.

Remember that for McGeer the normative commitments an agent concludes on by focusing on his reasons for acting in a certain way are not equivalent to the intentional states that he maintains. The agent's intentional states are "complex dispositions to think, speak, feel and otherwise operate in various mental and physical ways" (ibid, 90), and as such it is not always plausible to assume that the agent can instantiate such integrated behavioural patterns by only focusing on his reasons for acting in such a way. What the agent who attempts to integrate his various actions in a coherent whole expressing his self-understanding does is engage in a process of self-regulation[38]. As she argues, humans learn to regulate their behaviour in accordance to their normative self-ascriptions, or judgements, from engaging in common folk-psychological practices (see, for example, our first chapter). As such, these self-regulating practices are essential for acting individuals to learn to view themselves and others as competent, authoritative agents, to act as such and to hold themselves and others responsible for their actions.

On the one hand, in McGeer's framework the agent's first-person perspective is crucial because through it the agent develops a self-concept by making certain normative self-ascriptions that express the reasons the agent has for maintaining his intentional states. On the other hand, the agent uses these normative self-ascriptions

[38] See, e.g. Victoria Mc Geer and Philip Pettit, 2002, "The Self-Regulating Mind", *Language and Communication,* Vol.22, no.3, pp.281-299.

in order to regulate himself in appropriate ways by fitting his self-concept to his various displays of behaviour. And to do this, the agent needs to see himself as having certain empirical dispositions to act, think, and feel in certain ways which he ought to integrate in a coherent whole in order to act rationally and to be held responsible for his actions. That means that the agent needs to be able to adopt both a deliberative and a theoretical perspective on himself.

I find McGeer's considerations compelling and I think that if the account being developed in this thesis treats intentional states as dispositional then it will have to accommodate a certain practice of self-regulation in addition to the formation of a self-concept based on the agent's normative self-ascriptions. I don't find it plausible that by simply accepting certain reasons the agent can always be in a position to instantiate a coherent behavioural display which fits with the self-understanding formed by the acceptance of such reasons. I agree with Moran that in order to form a genuine self-concept the agent's ability to make up his mind and commit himself to a way of viewing the world over and above any kind of perceptual knowledge he has and inferences he makes is crucial, but I think that McGeer is right that in order to coherently express this self-understanding in his behaviour the agent will have to regulate himself in various ways. I also agree with McGeer and disagree with Moran that the means implemented by the agent in self-regulation can also be viewed as expressions of responsible, competent agency, even if they do not exclusively consist in focusing on one's reasons for maintaining one's intentions, beliefs, desires etc.

I'd argue that such self-regulation can be an expression of one's agency when it leads to the expression of one's genuine self-understanding in one's actions. For example, the gambler might make a resolution to stay away from gambling and subsequently implement certain means that would ensure that he maintains this resolution. In one case, he might make sure that all the dealers in his favourite casino hate him so that they refuse to deal with him whether he wants to gamble or not. This isn't the same as the gambler managing to stay away from gambling by taking a look at the blackjack table and thinking "no, gambling's bad for me". But it is still, I think, a case in which the gambler expresses his belief that gambling's bad for him and his intention to stay away from it no matter what, and so it is a case where he expresses his agency in his actions by acting in ways that express his self-understanding. If, in another case, the gambler is distracted by a friend just before he starts gambling again, I think that we would plausibly say that this doesn't consist in an expression of agency on his part because he didn't regulate his behaviour in any way in order to fit it with his normative self-ascriptions[39].

At least one important lesson to be had from Moran though is that over-reliance on a perspective from which one's mental states are viewed as empirical objects independent from one's active contribution can lead to one being alienated from one's own intentional states and mistakenly thinking that it is not in any way up to oneself to act in the ways one does. We then have to be very careful when providing an account of agency, in order to accommodate both the empirical and the deliberative stance. Also, there is a lot more to be said about the way an agent regulates his behaviour based on his understanding of himself as a competent, authoritative agent with certain intentional mental states. For example, we need to further clarify why an agent would attempt to produce a fit between his psychological understanding of himself and his actions.

The agentive drive for self-understanding

In order to further explore an account of agency which is based on the way the agent sees himself and the way such a self-understanding contributes to his self-regulation, I wish to examine Velleman's account of agency[40]. In Velleman's work, an acting individual's desire for self-understanding plays a crucial constituting role to his expressions of agency. This role consists in fitting the agent's self-conception, which consists of his self-attributions of intentional states, with the agent's actions, depending on his circumstances. By incorporating certain elements of this account in the view already emerging from the integration of Moran and McGeer's perspectives, I hope to end up with an account of the role of self-understanding in one's expressions of agency which respects both our nature as empirical fallible subjects who have to learn to regulate themselves in order to express their intentional states in their behaviour and the importance of our engagement with the world as authoritative, competent rational agents who can be held responsible for their actions.

[39] For a different example of a case in which externally manipulating one's behaviour can be understood as an expression of agency, see McGeer, 2007a, pp. 93-96.

[40] See Velleman 1992, 2009 and "The Self as Narrator".

Let's take a closer look first at what Velleman says about the drive towards self-understanding. Velleman introduces this motive in order to account for what plays the role of an agent in an acting individual's actions. The easy answer is that there is no one element in an acting individual that can be isolated and identified as the one that actively controls and is responsible for that individual's actions. In the case of intentional action, the organism that performs this action can be identified with the agent, since it is the one that brought this action into fruition. But things are not that simple. Velleman puts it best in the following passage:

> "Of course, the agent is a whole person, who is not strictly identical with any subset of the mental states and events that occur within him. But a complete person qualifies as an agent by virtue of performing some rather specific functions, and he can still lay claim to those functions even if they are performed, strictly speaking, by some proper part of him. When we say that a person digests his dinner or fights an infection, we don't mean to deny that these functions actually belong to some of his parts. A person is a fighter of infections and a digester of food in the sense that his parts involve infection-fighting and food-digesting systems. Similarly, a person may be an initiator of actions-and hence an agent- in the sense that there is an action-initiating system within him, a system that performs the functions in virtue of which he qualifies as an agent and which are ordinarily attributed to him in that capacity." (Velleman, 1992, pp. 475-476)

In order then to explain what plays the role of the agent in action we have to identify the functions characteristic of agency and explain what provides an acting individual with the capacity to perform these functions. The functions that characterize an agent's active participation in his behaviour are, for Velleman, the functions that constitute the agent's "acting in accordance to reasons" (ibid, p. 478). These functions have been taken by Velleman to include activities such as critical reflection on one's conflicting motives, endorsement and rejection of certain motives and the formation and implementation of intentions to act in a certain way[41].The intentions an agent forms have to be responsive to the reasons he has for acting, in this account of agency. And the agent's reasons for acting, in Velleman's view, depend on what he takes to be his motives for acting in a certain way under the specific circumstances he finds himself in. The agent's self-ascribed intentional

[41] See, for example, Velleman, 1992, p. 462: "The agent thus has at least two roles to play: he forms an intention under the influence of reasons for acting, and he produces behaviour pursuant to that intention".

states act as reasons for him to act in certain ways because, as Velleman argues, they enable him to understand his actions as being caused by motivating factors with which he identifies. When expressing his capacity for self-governed behaviour, the agent seeks to achieve "a folk-psychological understanding that traces the action to its causes in [his] motives, traits and other dispositions[.]"(Velleman, 2009, p.13)[42]. The main agentive function that we can derive from this account is the function of fitting one's self-attributions expressing one's intentional states with one's actions. And what enables the agent to perform this main function, in this account, is the agent's drive towards understanding himself as responding appropriately to his environment (in other words, as acting for reasons).

Why is such a motive what lies behind our capacity to function as self-governed individuals? Velleman postulates this agentive drive as the best solution to certain pressing problems that arise from attempts to reduce agency to event-causation. The main problem that Velleman encounters while attempting to provide such a reductive account is that in order to talk of the events that we can identify with expressions of agency, we must discover an element in the agent's cognitive organization that plays the role of the agent and that the agent cannot examine as something distinct from his own role as an agent. If this proves impossible, then it seems that if one still wants to provide an account of agent-causation then one would have to rely on the agent as an irreducible cause of action. Such a claim would be hard to defend, mainly because of the mystery associated with what the agent would be, in such a case[43]. Hence Velleman rejects this possibility and resorts to a reductive account of agency according to which the drive towards self-understanding is the only motive that can

[42] In his most recent work, Velleman specifies two modes of self-understanding that an agent can achieve. The first is the aforementioned folk-psychological understanding that the agent arrives at by viewing his actions as being caused by motivating factors that can be described in a folk-psychological framework (such as beliefs, desires and habits). The second kind of understanding is narrative understanding. As I understand Velleman, the agent achieves this kind of self-understanding when he behaves in ways that can provide him and others with a certain sort of emotional resolution. Velleman argues that the two kinds of understanding can combine in various ways in the agent's actions. An acting person can, for example, act in a way that is intelligible to him both as a product of psychological causal factors that he endorses and as a part of story that provides him with some kind of emotional resolution. For a discussion of these two distinct but interrelated modes of self-understanding, see Velleman, 2009, pp. 185-206. For the purposes of this chapter, I focus on folk-psychological self-understanding.

[43] For a recent defence of agent-causation, see Timothy O' Connor, 1995, "Agent Causation", in T. O'Connor (ed.) *Agents, Causes and Events: Essays on Indeterminism and Free Will*, New York, Oxford University Press, pp.173-200.

play the agent's functional role in a satisfying manner. That is because, as he argues, the agent cannot distinguish himself from his drive towards self-understanding without relinquishing his capacity for self-control. The agent cannot but be motivated by such a drive when he acts as such, even when he attempts to examine this particular motive. This motive is "functionally identical" (Velleman, 1992, p. 481) to the agent because the role it plays is the role that provides him with his status as such.

To better understand the reasons for adopting such a reductive account of agency, it will be instructive to discuss in more detail similar accounts whose flaws Velleman exposes[44]. The two theories Velleman examines for this purpose belong to Harry G. Frankfurt[45] and Gary Watson's[46], respectively. Frankfurt, in his search for the essential features of an agent's will to act, examines what he calls "first-order" and "second-order" desires that lead to action. In his view, first-order desires are the primary needs of an agent that prime him to act in a certain way and second-order desires are those that are fixed on the agent's primal needs and which either endorse or reject those needs. When an agent acts only based on his primal needs, then in Frankfurt's view he does not exercise his agency and his behaviour is not a product of rational reflection on his part on the way he acts. Nothing separates such an agent's actions from those of animals that only look to satisfying their primary needs without ever reflecting on whether they desire to satisfy those needs or not. But Frankfurt's second-order desires are in his account exemplary of such critical reflection and thus they are the main features of an agent's will to act.

These desires do not by themselves consist in the exercise of an agent's will to act in Frankfurt's account. What's missing is the transformation of these desires into what Frankfurt calls "volitions". Volitions for Frankfurt consist in the active endorsement by the agent of his second (or higher) order desires. If the agent actively endorses these desires then he turns them into volitions that motivate him to act in a certain way. So when an agent not only displays, but also endorses desires for or against the satisfaction of his primary needs or desires for or against the satisfaction

[44] For Velleman's discussion of these accounts, see his 1992, especially pp.470-480.

[45] See Harry G. Frankfurt, 1971, "Freedom of the Will and the Concept of a Person", *The Journal of Philosophy 68*, pp. 5-20.

[46] See Gary Watson, 1975, "Free Agency", *The Journal of Philosophy 72,* pp. 205-220.

of the desires corresponding to the agent's primary needs (such desires in Frankfurt's framework are of a higher order than the desires directly corresponding to an agent's primary needs) then the agent exercises his will to act. Frankfurt describes the agent's active endorsement of his desires as an act of identification on the part of the agent with his desires which leads him to act in a certain way. And this identification with higher order desires is essentially what turns an individual exhibiting a certain kind of behaviour into an agent exhibiting this kind of behaviour by exercising his will to act in a certain way.

Watson's account of what a person's agency consists in differs significantly from Frankfurt's account in that for Watson no amount of higher order volitions could be effectively posited as the agent's involvement in his behaviour. This is because as Watson argues a person's agency cannot be what chooses between higher order desires if it simultaneously consists of those desires. In other words, Watson's concern is that an agent cannot both identify himself with a second or higher order desire and also choose to reinforce or reject that desire. So a person's agency, a person's will to act, must consist in something other than second or higher order volitions. To that purpose Watson posits two different systems internal to the agent that play a role in his actions. He calls these systems the "motivational" and the "evaluational" system. On the one hand, the motivational system consists in the person's desires and wants that can lead him to action, i.e. in the person's motives. On the other hand, the evaluational system consists in the person's values, i.e. what would be worthwhile for him to achieve. In Watson's own words, "one's evaluational system may be considered one's standpoint, the point of view from which one judges the world" (Watson. 1975, p. 216). And the actions a person takes are in Watson's theory dependent on this person's beliefs and estimates of what would be the best course of action to take in any given circumstance.

As Watson makes clear these two systems can and do combine in order to lead an agent in behaving in a certain way (a person's valuing a course of action can combine with a person's desire to engage in that course of action). But in this account this is not always necessarily so, since a person's motivational system is more than capable to lead him to action without having to combine with his evaluational system. A person can indulge in his desires without ever evaluating them. So what can agency be reduced to in Watson's theory? Agency is in this theory constituted by the person's evaluational system and not by his motivational system. In Watson's view, Frankfurt is wrong to use higher order volitions as the building blocks of the will to act since such motives can always be separated from the exercise of one's will, from a person's agency, and thus they cannot be what constitute it. Watson believes he has solved this problem by positing his evaluational system, from which he believes the agent cannot ever entirely distance himself[47].

But this particular flaw that Watson sees in Frankfurt's theory and in response to which he posits the system of values which he believes constitutes a person's agency is a flaw that Velleman not only sees in Frankfurt's account but also in the account offered by Watson. Velleman argues that the agent can dissociate himself from both his higher order desires and his evaluational system and thus these features cannot be what constitute his agency Is Velleman right to attribute this flaw to both theories? In the case of Frankfurt's higher order volitions, both Velleman and Watson's arguments seem straightforward enough: the person acting can indeed at any time examine any of his higher order desires and renounce or embrace them respectively. So these desires cannot be reduced to the person's agency. But in the case of an agent's evaluational system, Watson does seem to have a good defence against Velleman's argument, as we have just seen. For Watson, an agent can dissociate himself from some of his values but he can never completely renounce his evaluational system without rejecting his identity as an agent.

Even though Watson's defence initially does seem compelling, I think Velleman's argument still runs through. And it runs through because what is at stake here and what Velleman's main concern is when attributing a common flaw to both Frankfurt and Watson's theories is not only whether an agent can partially or completely distance himself from the mental features that presumably make him an agent. It is also whether there is a principled reason of assuming that some of a person's mental features are functionally identical to his agency and some are not. In Watson's case, Velleman argues that even though an agent can embrace some values and reject others while retaining his evaluational system, there is no satisfying explanation of the reason some of the agent's values are part of the evaluational system embodying

[47] See Watson, 1975, p.216: "The important feature of one's evaluational system is that one cannot coherently dissociate oneself from it *in its entirety*".

his agency and some aren't[48]. So the problem here isn't only *whether* the agent can embrace or reject some of his values and include them in his evaluational system, but *what* this seemingly mysterious force which identifies or dissociates itself from such features consists in exactly.

So if Velleman is right, none of these accounts has coherently solved the problem of what exactly constitutes a person's agency, a person's involvement in his behaviour. The problem these accounts face is that they can't straightforwardly identify agency with higher order volitions or the system of values an agent uses since the agent can always distance himself from both of these elements without in turn losing his status as an agent. And how can an agent be identical to something which he can successfully conceive of as separate from his agency? Velleman's account has the benefit of overcoming this objection since we've noted, a motive for self-understanding is something that the agent is not in a position to distance himself from without losing his status as an agent.

Velleman also argues for the existence of this drive for self-understanding through examining a variety of experiments manipulating the social situations that their subjects participate in. Some of these experiments indicate that people tend to fit the way they see themselves, i.e. their self-conceptions, with their behaviour. Subjects have been found not only to fir their behaviour to the explicit claims they make, but also to manifest appropriate emotional responses when their situation is manipulated in certain corresponding ways[49].I'm not going to discuss these experiments and the complexities associated with their interpretation in this chapter, but I find it sufficient to note for now that such studies at least indicate the existence of possible empirical support for the idea that human beings frequently act under the influence of a motive to fit their self-conceptions to their behaviour.

[48] See especially Velleman, 1992, p. 472, footnote 26: "Of course, Watson refers not just to values lodged in the agent but to the agent's evaluational system; and he might argue that values are no longer integrated to that system once the agent becomes alienated from them. But in that case, Watson would simply be smuggling his concept of identification or association into his distinction between the agent's evaluational system and his other, unsystematized values. And just as Frankfurt faced the question how a volition becomes truly the agent's, Watson faces the question how a value becomes integrated into the agent's evaluational system".

[49] For Velleman's discussion of such experiments and their philosophical implications, see J. David Velleman, 2000, "From Self-Psychology to Moral Philosophy", *Philosophical Perspectives* 14, pp. 349-377 and Velleman, "The Self as Narrator", especially pp. 13-14. I explore these experiments and Velleman's take on them in more detail in Chapter 3.

Despite this, I don't think such empirical considerations are enough, by themselves, to demonstrate that this motive is constitutive of agency in the way Velleman argues that it is. For that, we have Velleman's argument that the attempts to reduce human agency to a state manifested when an acting individual performs the functions characteristic of self-governance seem to necessitate postulating the existence of such a motive. I do think that Velleman's considerations put considerable pressure on accounts of human agency and that if proponents of such accounts find this kind of constitutive drive a highly implausible possibility, then they'd either have to come up with a good alternative or resort to an account of non-reductive agent causation. I think that postulating such a motive can be helpful in providing an account of human agency, although as I hope will become evident from integrating this kind of account in the story emerging from McGeer and Moran's perspectives, such a motive can only be the driving force behind human agency when it operates within a specific kind of framework. This kind of framework is provided by the practices of folk psychology.

The common thread

The main reason Velleman's account is useful to get into at this point is that it can be applied to answering the following question: If we make certain (implicit or explicit) claims about ourselves that express our self-understanding, why is it exactly that we use these claims in regulating our behaviour appropriately? One answer to this has already been provided in our discussion of the normative role of folk psychology, as argued for by McGeer. It is because we take part in certain practices of explanation, prediction and evaluation of each other's behaviour that we learn to see ourselves and each other as being in a position to speak authoritatively both for our intentional states and our actions. We need to present ourselves as agents so that whoever attempts to communicate with us in a social setting will be able to anticipate our behaviour in order to respond in an appropriate manner. The people we interact with will also have to present themselves as authoritative self-regulated agents so that their behaviour becomes, in turn, something that we can anticipate and respond to.

This practice of self-presentation and reciprocating recognition of intentional action presupposes that the participants all share a common understanding. And this common understanding, as we have seen, is developed within a folk-psychological framework, constituted by an implicit theory of how the mind works. This implicit theory makes use of the concepts of intentional states such as beliefs and desires. The interacting participants act under a common understanding that both they and whoever they interact with know their own minds and what it is they're doing. Otherwise, as McGeer has argued, it would be hard to understand not only how these participants could interact in meaningful ways but also how they could engage in such purposeful, intelligible behaviour in the first place.

Velleman is useful in this context because of his account of a desire to make sense of one's actions performing the functions of agency and acting as a drive that motivates the acting person to integrate the self-attributions constituting his self-concept, his actions and his circumstances in a coherent whole. I am not sure if Velleman himself would agree with the contention that the central drive for self-understanding that constitutes agency is dependant on a folk-psychological framework, although he does talk of the self-understanding achieved by fitting one's self-conception (consisting of one's self-attributed intentional states) with one's actions as folk-psychological understanding. This does seem to presuppose that the agent needs to operate in a folk-psychological framework in order to regulate his behaviour in ways appropriate to his self-conception and his circumstances.

In any case, the argument here is that if human beings had not received training in folk-psychological practices from a young age, they would not have the motive to fit the claims they make about themselves with the various patterns they display in behaviour, expressing their self-understanding in their actions. That is because they would not understand what it means to act as self-controlled beings that are directly responsible for their states of mind and their actions and that can be taken as such by other such beings, in a reciprocal practice of interpretation, prediction and critical scrutiny.

In his (2009), Velleman seems to make throughout a similar point on the importance of social interactions for human beings' expressions of agency. He argues that we not only express the way we see ourselves in our behaviour in order to better understand ourselves, but that we act coherently in order to make it easier for other agents interacting with us to respond to us. This has as a consequence that their response is also better understood by us, and so on and so forth. Velleman uses this idea to develop an account of how it is we share a lot of the same moral values, since they depend on certain aspects of our human nature that we all share and we all express in our behaviour. In general then, I don't think he would be unsympathetic to the account developed in this thesis. One difference that I think exists between his account and the one based on McGeer's viewpoint is this: Velleman seems to think that the desire for self-understanding leads to the specific way that the social interactions between self-controlled individuals take place. Conversely, in our own account social interaction within a folk-psychological framework provides the guidelines for the specific regulative functions that the agents should perform in order to understand themselves. It's hard to see, after all, how an agent could regulate himself appropriately in order for his actions to be explained by reference to his intentional states if he did not already have an understanding of how these intentional states are manifested in human behaviour.

This also seems to have the consequence that a human being that was not part of a society depending on such practices would not have the capacity to regulate its behaviour as an authoritative agent. Such a conclusion would seem to naturally fit an account such as McGeer's in which self-regulation and first-person authority are acquired capacities. In short, I don't think the motive to make sense of one's actions could be constitutive of human agency if it didn't operate within a conceptual framework consisting in a folk-psychological implicit understanding of the causes and effects of human behaviour. A motive to understand one's self and one's actions within the environment one acts in may very well be an innate human drive, as Velleman seems to believe[50]. But I would add that such a drive can enable acting individuals to perform functions such as regulating their actions in order to instantiate their self-attributed intentional states in their behaviour only if the individuals motivated by that drive made these self-attributions in the context of a shared understanding of what it means to have intentional states that express one's

[50] See e.g. Velleman, 2009, p.17:
"Anyone who has dealt at close quarters with infants or toddlers knows that the human animal is born with a voracious cognitive appetite.
During its second year, the child acquires a conception of itself as cognizable object, a thing to be understood. And then it comes to see that understanding this particular thing is quite different from understanding any of the others…[t]he inquirer learns that he can make sense *of* himself, as object, by making sense *to* himself, as subject-that is, by doing what makes sense to him."

active participation in one's behaviour. The idea here is that if a human being didn't have a self-concept consisting of self-attributions of intentional states that enabled it to recognize the behavioural patterns constituting these states and the effects the instantiation of such patterns would have on the behaviour of similar beings, then it would not be able to instantiate such patterns through actively shaping its various responses to its environment.

In summary, this is the proposed account of how human beings like us act as self-regulated agents who have authoritative knowledge of their own minds and control over their actions. The insight we got from Moran and McGeer's viewpoints is that our first-person perspective on our behaviour and our environment enables us to adopt certain self-attributed commitments to behave in a variety of ways expressing our intentional states. We are responsible for such commitments because they are responsive to the reasons we have for acting in certain ways and these reasons depend on how we respond to the circumstances we find ourselves in. These commitments don't depend on an introspective capacity that enables us to perceptually recognize our cognitive states. Instead, they are formed through our taking a stand to what our engagement with our circumstances commits us to. For example, a commitment of ours to the statement that a glass of water is poisonous could be formulated through our discovery that drinking from this glass causes intense pain. The formation of such commitments through our active engagement with the world constitutes our expressions of mental agency.

But the question of what exactly these commitments amount to ends up revealing an important difference between McGeer and Moran's perspectives. For Moran, such active commitments constitute our intentional states. If we were to bring about our intentional states in any other way other than by focusing on our reasons for expressing them in our behaviour, that would indicate, according to him, that we are alienated from these states. So for Moran our active normative commitments are our intentional states. I have chosen to disagree with Moran and side with McGeer on this issue, since I believe that her account makes for a more plausible explanation of how it is that empirically fallible beings like us express their intentional states in their behaviour. In this view our normative commitments constitute certain judgments we make on how we should respond to our environment. Since our intentional states, in this account, are complex dispositions to act, think and feel in various ways, our judgments do not always automatically result in the instantiation of these states. Instead, we frequently engage in a process of self-regulation in order to develop and reliably express these dispositions in our behaviour. This kind of self-regulation has two features: First, it can enable us to make use of indirect means in order to exercise our agency in our behaviour. Secondly, it has been developed through our engagement with each other while sharing a common understanding of what it means to express one's intentional states in one's behaviour, in short, what it means to act as an agent.

In considering this account, we were also led to the following question: Can we elaborate on the nature of human beings' ability to play the role of the agent by fitting their normative self-ascriptions to their behaviour? By using Velleman's perspective, the answer we ended up with is this: the functional role of the agent is played by our drive to make our actions intelligible in order to act in a coherent manner and to efficiently communicate with others within a social setting. This setting presupposes that we have the capacity to know our own minds and to be in control of our actions. The drive towards self-understanding is what lies behind our self-regulative capacities and hence, what lies behind our expressions of agency in our behaviour. Even though the drive towards self-understanding might be an inherent aspect of our nature, it only develops as the motive that is constitutive of our self-regulative capacities within a social setting in which the interacting participants have a common understanding of what it means to be a self-regulated creature and of the ways in which such a creature would express its intentional states in its behaviour[51]. And our best such understanding, as also argued by McGeer, is engendered in our folk-psychological practices of interpreting, predicting and criticizing each other's behaviour.

[51] Perhaps another way of putting this is that the inherent desire for self-understanding evolves into a desire for coherent self-presentation, to oneself and to observers of one's behaviour, within such a social collaborative setting.

Chapter 3

From Self-Regulation to Human Agency

Introduction

In this chapter, I intend to further develop our account of agency, i.e. of how creatures like us frequently can engage in purposeful behaviour so that we are active in shaping our actions in ways expressing our unified points of view (or ourselves), by examining the differences between purely self-organizing and self-regulating intentional systems and arguing that human beings fall under the latter category. I will start by examining Daniel Dennett's observations on self-organization and the role of a self-concept in human behaviour[52]. I will then supplement these observations with Jenann Ismael's view that we are self-regulating creatures because we have developed, on top of our self-organizing substructure, a kind of virtual map of ourselves as moving through our environment and we use this map (which acts as a locus of unified information and control) in order to regulate our behaviour in various ways[53]. Even though I think that Ismael's work can provide a framework in which both the differences between self-organization and self-governance and the fact that human beings belong to the latter category can be appreciated, it is still not clear, in this view, how exactly our conception of ourselves and our environment influences our behaviour.

Because of this vagueness, I intend to use J. David Velleman's account of human agency in order to argue that we are motivated to fit our self-concepts to our actions because of our drive to make our actions intelligible to ourselves and others[54]. Velleman argues that this drive has the function of forming coherent narratives from our self-concepts, actions and circumstances. I argue that Velleman's story is useful for our account, as long as we understand the motive for self-understanding not as a

[52] See Daniel C. Dennett, 1992, "The Self as a Center of Narrative Gravity", in F.Kessel, P.Cole and D.Johnson (eds.) *Self and Consciousness: Multiple Perspectives,* Hillsdale, NJ: Erlbaum, Dennett, 1993, *Consciousness Explained,* London: Penguin Books and Dennett, 1996, *Kinds of Minds,* New York: Basic Books.

[53] See Jenann Ismael, 2006, "Saving the Baby: Dennett on Autobiography, Agency and the Self", *Philosophical Psychology* 19(3), pp. 345-360 and Ismael, Forthcoming, "Selves and Self-organization", available at http://homepage.mac.com/centre.for.time/ismael/

[54] See J. David Velleman, 1992, "What Happens When Someone Acts?", *Mind* 101, pp. 461-481, Velleman, 2009, *How We Get Along*, Cambridge University Press and Velleman "The Self as Narrator", available at https://pantherfile.uwm.edu/hinchman/www/Velleman-Dennett.pdf

central controlling module but as a motive that enables us, as a whole, to engage in actions that express our self-concept in an intelligible manner.

I conclude with the claim that seeing ourselves as being self-regulated in this way is a promising step towards giving an account of human agency. That is because it enables us to understand how it is that agents actively express their point of view in their behaviour in order to engage in purposeful behaviour. In this account, the agents' point of view is constituted by their self-concept (which is a concept of themselves as moving through their environment) and it is expressed in their actions because of their motive to make their actions cohere with the way they see themselves. Be that as it may, my argument is that more work has to be done in order to give a full account of human agency, by turning to the fact that human agents are a part of a complex social network in which they not only have to act as self-regulated individuals, but they also have to interact with others who act as such.

Self-organization as a first step towards human agency

While observing the behaviour of a human being, one notices certain distinct patterns in the movements on display. Such patterns may consist in the repeat performance of certain simple or complex movements under a certain time frame or in the coordination of different kinds of behaviour in one single behavioural expression. While painting a wall one moves a brush in wide continuous strokes. A dancer moves his body in the rhythm of the music by integrating the different movements of his various body parts into one continuous behavioural expression. A swimmer performs repeated arm and leg movements and positions her torso in a particular way so as to maximize her swimming speed while manifesting a rehearsed swimming technique. A speaker in a seminar moves his arms in various ways reflecting the tone and rhythm of his speech and conveying the meaning of his words. A teacher writes on a blackboard, while at the same time keeping an eye on her pupils and making sure that they are paying attention to what she writes. The verbal reports of the questioned subjects also fit in certain behavioural patterns. The swimmer, when asked why she was moving the way she did, will plausibly answer that she was trying to manifest a certain swimming technique in the most efficient manner possible in order for her to win the race she finds herself in. The speaker in the seminar might say that he was trying to convey certain meanings and that his hand movements not only help him convey these meanings but they also help him concentrate on his speech. And so on with the rest of the aforementioned examples.

Such perceived patterns in behaviour seem to indicate that there is meaning to these kinds of movements, that they occur in a specific way because they express purposeful, unified, goal-directed behaviour and thus that they are not just random moves but actions of intelligent agents who can be said to know their own minds and to be able to express themselves in their actions in order to achieve certain goals. Being such an agent that engages in purposeful behaviour also seems to imply that such a creature is active in shaping its actions in certain ways and hence that it is responsible for them, in a way that a creature compelled to behave in certain ways due to various external factors, isn't.

Can we make sense of such an appearance of agency in creatures like us by giving an account of how it is that we engage in such purposeful behaviour? If these patterns do indicate purposeful behaviour and if they correspond to specific self-controlled actions of human agents, then these agents seem to have the ability to specify and execute the various parameters associated with these actions. In the case of the swimmer taking part in a race, we can imagine that there are a number of different parameters specified in her action of swimming in freestyle. These parameters can be viewed as encompassing everything from the specific position of her limbs (e.g. her arms must be relaxed below the elbow so as to make it easier for her to swiftly move them above her head and into the water) to the specific movements of her body through the water (e.g. her hands must not wander too far from her body while stroking the water, she must avoid splashing the water as much as possible, instead letting her arms smoothly glide in it while she draws short gasps of air, tilting her head sideways at the right times). All these complex movements culminate in the single action of freestyle swimming. It is hard to imagine how such an action can occur without the aforementioned parameters being specified somehow by the agent and being applied in the continuous motion of her body. But if these parameters are specified and applied in action, then what is doing the specification and application? Is it plausible to say that this would be a central executive entity controlling the actions of the agent in question, and that this controller is located somewhere in the agent's brain? Is there a way to answer such a question by observing the swimmer's behaviour, which also includes the swimmer's verbal reports?

If asked why she was using the technique of freestyle swimming, the swimmer would presumably answer that she believed that she had to adhere to the rules of the race she was in so that she wouldn't be disqualified. She might also answer that she wanted to win that race and therefore utilized the freestyle technique in the most efficient manner possible, based on her previous rehearsals of this technique in the years she's been practicing it. If questioned some more about the specifics of her action, she might talk about her sudden realization during the race that one of her opponents was gaining on her, her fear of losing the race and her attempt to concentrate on her swimming technique while not letting these other thoughts interfere with her actions.

Who was the one having these desires and beliefs, the one reflecting on what was going on during the race and attempting to withhold negative thoughts from interfering with the main act of manifesting a swimming technique in the most efficient way possible? The swimmer, of course. But where in the swimmer did all the aforementioned processes occur? Shouldn't there be a certain area in the swimmer's brain where beliefs and desires are being expressed while being allowed (or not) to play a role in her actions? Or where the swimmer's reflection on her actions, her feelings about these actions and her memories of previous actions are being examined and attempts on blocking or allowing their interference in the swimmer's current actions are executed? It seems that this line of thought leads to positing of a self in the swimmer that is the bearer of reflections, feelings, and memories and of the swimmer's beliefs and desires. This self might be located somewhere in her brain and control all her purposeful actions such as her verbal reports and her exercising a swimming technique in order to win the race she's in. Furthermore, the swimmer's behaviour is purposeful and constitutes the actions of an active agent because of the self's controlling influence

Unfortunately, this line of thought is very familiar, and it leads to a pretty familiar objection as well. It's Daniel Dennett's refutation of what he calls Cartesian materialism and the illusion of the Cartesian theatre[55]. The Cartesian theatre, as has

[55] See especially Dennett, 1993.

been described by Dennett, is the conception of the self as an entity residing in an acting agent's brain, examining his various perceptions, thoughts and feelings and exercising executive control on the agent's actions. In its radical form, this conception posits the self as something distinct from the body in which it resides but which controls the body nonetheless. This sort of dualism is implausible, but this isn't the only form of the Cartesian theatre conception that can be found in the relevant literature. As Dennett takes great pains to show, this kind of view of the self is pervasive even in materialistic views of the relation between the mind and the body.

This pervasive view of the self as centrally located in a specific place in the brain is what he calls Cartesian materialism. In such conceptions, the self might not be viewed as a distinct entity independent from the body it controls, but it is still viewed as something which is located in the brain and which has to apply executive control to all the intentional actions of the agent of which self it is. The self in such views can be anything from a number of distinct cells in the brain to a specific functional organization of physical properties. As Dennett aptly argues, positing a mini-self somewhere within the acting individual as representing a distinct area in which all input from the environment has to be processed before it leads to behavioural output is highly problematic. I find his rejection of views positing a controlling central area in the brain compelling since my view, at least in the case of agency, is that we should avoid arguing that an acting individual acts in a self-controlled manner because it has a mini-agent within it guiding its actions. Instead, I think that the whole of the acting individual should be seen as responsible and in control of the actions it performed as an agent and not just a part of that individual (a part that presumably makes all the important decisions by itself).

I agree that such Cartesian views postulating an elusive executive entity should be rejected. What alternatives do we have for explaining seemingly purposeful, unified behaviour, if not as being the product of a controlling entity within the acting intentional system? Dennett argues that we should understand this kind of behaviour as the product of self-organization. Self-organization is the process that enables complex intentional systems to give the appearance of engaging in unified, goal-driven behaviour while not being in fact guided by any kind of controlling entity within their substructure or any kind of locus of unified information and control in general. The classic example that Dennett gives of a self-organizing system entails that of an ant colony. The ant colony as a whole can be understood as a self-organizing intentional system whose basic constituents, the ants, collectively engage in all kinds of purposeful behaviour, such as the formation of groups that overwhelm other insects invading their nest. Naively, we can provide a lot of explanations of such behaviour that point towards the existence of something in the colony influencing the ants. Maybe they are following their queen's commands or some kind of general plan they have devised in case they have to protect their nest. But knowing ants, we know that no such explanation would be accurate. Instead, ants just follow their innate predispositions and as such, act individually. But because they behave as such in tandem with other ants behaving in similar ways, an overall appearance of purposeful behaviour is achieved. The ants' innate dispositions are constrained by each other's behaviour and by the ants' environment. The ants' interaction with each other and with their environment constitutes the general framework in which this behaviour is taking place, in a way that allows them to engage in collective displays of such behaviour. As Dennett puts it when describing such a colony,"[w]e now understand that its organization is the result of a million semi-independent little agents, each itself an automaton, doing its thing." (Dennett, 1993, p. 413)

How does Dennett extend this case to the case of human beings? According to him, we are similar to the ant colony in that we are also self-organizing intentional systems. We are made up by a multitude of content-relaying processes which are autonomous in a similar way to which the ants in the colony are autonomous. The content of these processes consists in various kinds of information received from the environment. The individual, distributed sub-mechanisms that process the various input received by the acting individual compete and cooperate with each other, with the result that the content they carry becomes more or less influential for the individual's behaviour. The most prevalent kinds of content at a given time might get to influence the individual's behaviour (which might in some cases include the individual's verbal reports) and be retained at the individual's memory. The extent to which any or all of these effects take place at any given time depends on how influential content-relaying processes carrying information that is relevant to these effects become.

This is the basis of Dennett's answer to the Cartesian theatre. There is no central controlling self since the agent is made up by autonomous subsystems carrying different kinds of content. In such a case we'll avoid the problem of having to posit a mini-agent that does all the work. But I believe that accounting for human agency requires more than just a simple analogy to the ant colony. It seems that we have the ability to form long-term goals and engage in actions expressing our unified point of view in a way that requires more than pure self-organization, in other words that we can act as agents in a way that systems like the ant colony cannot[56]. In order to explain why that might be the case, I think we first need to understand that the difference between us and purely self-organizing systems like the ant colony is not only quantitative, but also a difference in kind. It is not simply a difference of the degree of complexity but also a difference in the distinct capacities that an increase in system complexity would entail.

Dennett himself seems to agree with this assumption, since he does talk about an ability human beings have that other self-organizing systems don't. That is the ability to "grow self-representations." (ibid, p.430) A self-representation, according to Dennett, "plays a singularly important role in the ongoing cognitive economy of [a] living body, because, of all the things in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself." (ibid, p. 427) More specifically, the agent's autonomous information-processing subsystems function as a "Joycean Machine", in Dennett's framework, because they can integrate a variety of information in the form of a narrative stream organized around the system's concept of itself. The model of the system's self consists of the system's self-attributions of mental and physical characteristics, which are based on its self-interpretation of its behaviour[57].

[56] Dennett himself seems to share this intuition. See e.g. ibid, p.228:" We are *not* like drifting ships with brawling crews; we do quite well not just staying clear of shoals and other dangers, but planning campaigns, correcting tactical errors, recognizing subtle harbingers of opportunity, and controlling huge projects that unfold over months or years".

[57] See e.g. ibid, pp.428-429: "An advanced agent must build practices for keeping track of both its bodily and "mental" circumstances. In human beings, as we have seen, those practices involve incessant bouts of storytelling, some of it factual and some of it fictional…….Thus do we build up a defining story about ourselves, organized around a basic blip of self-representation".

Is such an informational stream that is built around the system's self-concept used in guiding the system's actions? Dennett seems to be of two minds on this issue. On the one hand, as we have seen, he does talk of the self-concept as a crucial element for the system's actions, but on the other hand, he frequently refers to the self, as represented by the system, as an abstract that is postulated for the purposes of explaining and predicting the system's behaviour without being used by the system itself in guiding its behaviour. I take it that one reason that Dennett has for avoiding talking too much about the self-concept's influence on action is that it would threaten to take us back to the problematic conception of the self as an invisible integrator of information who controls the agent's actions. But there might be a way to avoid relapsing in such a fashion while still preserving the importance of the fact that some highly complex intentional systems can construct a model of themselves that they use in guiding some of their actions. Jenann Ismael is an author that shares this assumption, whose theoretical framework will help us better understand the role played by an intentional system's self-representation[58].

Self-regulation as distinct from pure self-organization

Ismael examines Dennett's views on self-organization and concludes, essentially in agreement with what I have said so far, that there are inconsistencies in Dennett's work that need to be resolved. To do that we must distinguish, she argues, between intentional systems like the ant colony that engage in seemingly purposeful behaviour only as a result of their self-organizing substructure and intentional systems that can represent themselves and organize some of the information they receive from their environment around their concept of themselves. These systems can then genuinely engage in purposeful behaviour that is relevant, for example, to their long-term goals because they are guided by the content in the informational stream centred on their self-concepts.

In these systems some of the information received by their environment through their senses can be diverted and integrated in a unified informational stream, which is used by them in guiding their actions. This stream, which Ismael calls the "Joycean monologue", functions as a kind of virtual map containing a representation of the

[58] See Jenann Ismael, 2006,"Saving the Baby: Dennett on Autobiography, Agency and the Self", *Philosophical Psychology*, vol.19, no.3, pp.345-360 and "Selves and Self-Organization", *Minds and Machines,* forthcoming.

world the system finds itself in, as centred on that system, and a model of the system itself which contains information relevant to navigating its environment (features from the system's psychological and physical constitution, such as its personal history, its intentional states and parts of its body)[59].The information contained within the Joycean monologue makes it possible for the intentional system to engage in processes such as deliberation and long term planning, processes which lead to actions that are rational because they are guided by them. In Ismael's account then an intentional system functions as a rational agent because of the Joycean monologue's influence on its actions.

What are the major differences of such self-governing systems and purely self-organizing systems? According to Ismael a self-governing system is more flexible when it comes to its responses to its environment. In other words, self-governing systems have more possibilities for action than self-organizing ones. Ismael argues that this is because self-organizing systems display purposeful behaviour only as a result of their components' innate dispositions and they always respond in the same way to the same environmental input. In contrast, a system that represents itself and its environment and uses these representations to guide its actions can respond in different ways to the same input, depending on the information contained in its self-representation and the goals it is pursuing at the time of action. A self-organizing system can learn to respond in different ways to the same input but this happens only as a result of conditioning, which takes more time than responding in the flexible manner exhibited by the self-governing systems.

To further develop this point, in accordance to Ismael's observations, systems which only rely on self-organization are, in general, faster and more efficient than self-governing systems. But self-governing systems have a greater flexibility when it comes to adjusting their responses to their environment. Human beings are different than ant colonies in that they do not necessarily need external conditioning in order to respond in novel ways to the same external stimuli. They respond in a novel way because they have different goals in every given circumstance depending on their evolving representation of themselves. Increased internal complexity in this case

[59] See Ismael, 2006, p.350: "We don't just monitor our spatial locations, we keep track of our physical properties and our representational states, described in explicitly intentional terms, and we incorporate all of it into our self-models, together with an explicit record of our personal histories."

enables such intentional systems to process the same external input in different ways and subsequently to respond in different ways to the information they receive from their environment. Hence the range of possible actions in self-governing systems seems to be far greater than the one in self-organizing systems. Self-organizing systems display coordinated behaviour because of higher-level constraints arising from the interaction of lower-level individual components and hence their options for action are limited by these higher-order constraints. In contrast, self-governing systems which form representations of themselves and their environment which they use as a kind of virtual map have increased options for action depending on the complexity of these self-centred representations.

The obvious upside here for self-governing systems, as Ismael notes, is their ease of adaptability to a range of different environments and to potential rapid change in their environmental conditions. The downside is that their responses to input from their environment will usually not be as fast and reliable as those of self-organizing intentional systems which operate in optimal conditions. Such conditions for self-organizing systems will be those that presumably best utilize the innate tendencies of the elements constituting these systems. But human beings plausibly do not always or even frequently operate under such conditions and they usually have to adapt to a lot of rapid changes in their environment. Mere evolutionary conditioning does not seem enough for the efficient function of such systems and this is why self-regulation could be a valuable tool for them. Self-regulation is faster than evolutionary conditioning in adjusting a system's responses to its environment and it adds greater flexibility to the ways an intentional system responds to a rapidly changing environment.

Note here that self-regulation allows for more efficient and rapid adjustment of the way self-regulated systems respond to the input received from their environment, but not for more efficient and rapid behaviour triggered by certain input. Ismael makes this point explicitly in her following remarks:

> "[S]elf-governance involves a real departure from self-organization and brings with it genuinely new capacities. It brings the sort of flexibility that allows not just quick response, but immediate adaptation of stimulus-response connections to wider circumstance………The claim is not that self-governors will adapt behaviour more finely to stimuli, but that they will change their *response functions* with changes to stimuli. Reflexive responses are excellent

> (much better than self-governors under many conditions) at guiding rapid motor behaviour………What they are *not* is flexible at the level of response function."(Ismael, forthcoming, p.13)

In fact, this is one of the differences motivating Ismael to argue that on the one hand intentional systems with self-regulating capacities are indeed different from self-organizing systems without such capacities, but on the other hand that the one does not exclude the other. It would be implausible to argue that a system could be self-regulating without it having developed from a self-organizing structure that is also used in carrying out the system's actions even when these actions are being shaped through the exercise of self-regulation. The self-concept is not used in micromanaging every single detail of the motor functions that are part of the system's action, since such functions are controlled far more efficiently by the system's various autonomous subsystems. But self-regulation is crucial when it comes to the overall way a system responds to its environment, because the kind of responses the system tends to exhibit depending on the stimuli it receives will be different as the self-concept evolves and adapts to the system's varying circumstances.

Based on Ismael's considerations, I think we can understand to a greater degree the ways in which self-regulating intentional systems are distinct from purely self-organizing ones, in addition to the view that these two types of systems are not mutually exclusive. Self-regulating intentional systems are different from purely self-organizing ones for two main reasons. One is that self-regulating intentional systems have the capacity to construct a model of themselves and use this evolving representation in order to navigate their environment. As Ismael notes, this doesn't mean that everything these systems do accords to the information relating to their self-concept. It just means that there are certain kinds of behaviour, such as behaviour that depends on long-term goals that the system has, that require the regulating influence of an explicit model of the system. Secondly, self-regulating systems demonstrate a greater flexibility to adapting the way they respond to their environment than purely self-organizing systems, even though they still depend on autonomous subsystems controlling motor function in a much faster way than would be possible by self-regulation alone.

The self-concept as the locus of information and control: who's in charge?

It seems that by relying on human beings having the ability to regulate their behaviour according to their self-concept we can provide the basis for accounting for human agency. We sometimes act in a self-knowing, self-controlled manner because we act according to the way we view ourselves. Furthermore, because the way we act in those circumstances expresses our point of view, we can be held responsible for our actions. Hence, the ability to regulate one's actions according to one's self-attributions seems to justify our intuition that we can act as agents and exert some kind of active influence in our behaviour.

Not so fast. I think the way self-regulation has been set up so far, even though it can be granted that it's distinct from pure self-organization, still seems too vague. One significant problem is this: If it is granted that a highly complex intentional system can represent itself as moving through its environment and create an informational stream centred on this representation, how exactly does this self-concept (and all the relevant information built around it) act as locus of information and control for the system? How is it that the system's self-centred informational stream can exert any kind of influence on the system's actions? Focusing a little more in detail on Ismael's story might help with this problem.

As we have seen, in Ismael's story self-regulating systems form a kind of virtual map (what she calls the Joycean Monologue) containing a model of the system's self (including self-attributions of states such as beliefs, desires and intentions) and a representation of the environment that this system is interacting with. Ismael's ship metaphor might clarify a bit how self-regulation works. Suppose you take a ship having to navigate its environment by using a map of its location, Ismael says. This is a map that is constantly updated by the ship's instruments every time the ship receives new readings from its environment, readings which include the ship's own movements, plotted course etc. So the map not only contains information about the ship's environment, but also a model of the ship itself used to navigate this environment. Furthermore, there does not have to be any intelligent captain using the map in the ship. The map is formed by the ship's own subsystems (which are meant to be analogous to the autonomous subsystems in self-organization) and it is also used in guiding the ship's course by those subsystems.

More specifically, Ismael presents her view of how the ship regulates its behaviour thusly:

> "If we focus just on the evolving contents of the map, ignoring all of the activity that's not explicitly represented there, what we see is an informational stream whose content is that of an evolving, objective representation of the spatial landscape centered on the ship. This informational stream receives input from the environment in the form of informational states that have the contents of self-locating beliefs, but it is propelled by an internal logic that transforms those states into prescriptions for action, and that internal logic has roughly the form of deliberation. Those prescriptions for action, moreover, feed back into the ship and, provided all goes well with the rest of the machinery, guide the movements of the ship."(Ismael, 2006, pp. 349-350)

The internal logic the ship is propelled by is the interesting part here. What exactly motivates the ship to use the information contained in the map consisting of a model of the ship and its environment in order to guide its movements? How does deliberation take place within the map? If we just leave the description of self-regulation at that then it becomes very mysterious why self-regulating systems use their self-concepts in order to coordinate their actions. How exactly does the informational stream revolving around these self-concepts influence the systems' actions? It seems that if we are not careful, we might fall back to talking as if the self-concept itself, or the virtual map itself, exerts some kind of active influence on the system's behaviour. Ismael herself seems to occasionally fall back on talking this way:

> "The role that this informational stream is playing is something like that of the CEO of a vast, and largely self-regulating bureaucracy: unaware of the day to day activities that keep the system running, but setting long term goals, keeping track of the system's progress, and exerting influence needed to nudge behavior in the direction of goals."(ibid, 350)

This makes it sound as if the informational stream itself, or the self-concept, or the virtual map, or however else you choose to view the unified locus of information and control used by self-regulating systems, does all the work relevant to the system's expressions of agency. The self-concept deliberates, it sets long-term goals, it reflects on certain motives and decides which motives to reject and which to reinforce, and it motivates the system to engage in all kinds of purposeful behaviour. But surely this can’t be right? Especially seeing as we’ve started with the assumption that we need to reject the conception of a mini-agent hiding within the system’s brain, to whom we trace all the system’s actions, when it acts in a self-controlled manner. So what we should say is that the informational stream doesn’t actively do anything. In other words, the Joycean Monologue itself does not exert any kind of active influence on the system’s self-organizing substructure. It is the autonomous subsystems that do all the work, in ways extensively described mainly by Dennett but also by Ismael, and they sometimes also use the information contained in the Joycean Monologue, in cases, for example, where the acting individual engages in long-term projects.

I’m fairly certain that both Ismael and Dennett share this view. They both want the self-organizing substructure, Dennett’s Joycean Machine, to do all the work in producing the acting system’s behaviour. The Joycean Monologue might be a unified locus of information and control which enables the acting individual to act as an agent, but that is because it is used as a reference point by the Joycean Machine which does all the relevant causal work. The question here is this: Can we account for how the Joycean Machine uses that information, without falling back to the Cartesian Theatre?

Distributed agency

The reason I take a view favouring distributed agency as opposed to “confined” agency should be spelled out. The agent is not an executive mechanism which makes all the important decisions pertaining to self-control. Instead, the whole person should be viewed as an agent. Why is that? First of all, as we’ve seen in our discussion of Dennett’s rejection of the Cartesian theatre, it seems that providing a plausible picture of the factors enabling a complex organism to act in a self-controlled and purposeful manner entails opposing a view in which a single element in the organism controls its actions. Taking Dennett’s Multiple Drafts Model, the organism is made up of a variously interacting multitude of information-relaying processes. Making a single element within this complex hierarchy the one which actively leads to the organism’s actions that express its agency seems to mean that all the content that is used in these actions would have to somehow first be processed by this crucial cognitive element. Hence, the controlling mini-agent is postulated. If the processes manipulating the kinds of informational input received by the organism are not all located in one single spot within the organism but are instead widely distributed in its body (and perhaps in its environment if one accepts extended cognition views), then the mini-agent would have to be able to somehow examine all of the relevant content before it decides which kinds of content should lead to the organism's purposeful actions.

But as Dennett notes, this makes for a highly improbable and in all likelihood empirically untenable view. Focusing on the brain alone, input received by the organism is processed at different times and places within its cognitive architecture. The content received by visual means might interact with content received by auditory means in order to lead to a certain kind of behaviour. Whether this behaviour expresses agency or not should not depend on both kinds of content being manipulated by a single kind of entity first. Even in this simple example, is it plausible there would be one confined area within an organism's cognitive architecture in which both kinds of content made an impact? Even if this looks plausible, the amount of plausibility of such scenarios decreases the more the complexity of the means through which the acting organism receives informational content from its environment increases. The more distributed cognition is seen to be, the less plausible a mini-agent becomes. In the case of human agents, the complexity of the means in which they can acquire information from their environment and the depth of their cognitive architecture indicate that a view of highly distributed cognition is closer to the mark than a view in which all the important functions occur in a confined area. To accommodate such a distribution of content within an account of agency, I also think, following Dennett, Ismael and other defenders of distributed cognition[60], that we should distance ourselves from the idea that there is a mini-agent within the acting person that expresses its agency in the person's actions. Instead, I think an account of agency should have as one of its conclusions that the acting person is the agent, since this person is made up of complex distributed processes that lead, one way or the other, to the kinds of behaviour this individual displays.

[60] See, for example, Andy Clark, 2007, "Soft Selves and Ecological Control", in D. Ross, D. Spurrett, H. Kincaid and G.L. Stephens (eds.), *Distributed Cognition and the Will,* MIT, pp. 101-122.

As it is, accepting the idea that the agent is, in a sense, a very complex entity brings with it a new host of problems. Agency that is simple and confined to some sort of executive mechanism or an area wherein executive functions take place has certain advantages. The most important one is that we can answer questions such as "who is responsible for that action?" and "whose goals are being expressed in action?" by pointing to the single executive entity or to the area in which executive decisions relating to processes such as long-term planning, goal-setting and deliberation take place. Not only does it seem that we can straightforwardly answer questions having to do with responsibility and self-control, but we can account for how a unified behaviour is being produced by a complex entity. The unity of its actions is derived from the unity of the "mini-agent" within it producing them. Even though the problems with such a view, as we have seen, make such an account of agency untenable, an account that respects the acting individual's cognitive complexity should also respect the unity of its goals and its purposeful behaviour. Eliminating the unified agent whose goals and intentional states are expressed in his actions seems to me as equally implausible as postulating such a unified agent as a simple executive entity to be found in a single area within the acting individual's cognitive structure.

Narrative control

J. David Velleman has written extensively on the topic of agent causation and on the way a self-concept can be used in action[61]. As we have seen in the previous chapter, he reduces the agent to the acting individual's desire for self-consistency, which enables the agent to perform the basic functions of agency. The acting individual acts as an agent because it is motivated to fit the way it sees itself to its actions so that its actions make sense as expressing its viewpoint. The agent expresses his viewpoint in his actions because he acts for certain reasons. The agent's actions express his reasons for acting when they express his intentional states and the states that express the agent are the ones that constitute his self-concept. In short, the agent forms a self-concept which expresses his judgments of how things are in his environment and these judgments express his intentional states, which in turn can be expressed in the agent's actions.

[61] See especially Velleman, 1992 and Velleman, "The Self as Narrator."

For our present purposes, what I think can be usefully applied to our present discussion is that this desire for self-consistency drives the acting individual, in Velleman's account, to form a coherent narrative consistent with its self-concept, its actions and its environment. Velleman has talked about the "self as narrator", which is his idea that the motive to make sense of one's actions acts a narrative module that fits the agent's self-attributions to his behaviour. One of the points emphasized by Velleman is that one's self-concept is not only formed as a response to one's behaviour but it also feeds back into this behaviour influencing one's actions accordingly. As such, there is a circular causal relation from action to the self-concept and from the self-concept to action.

This interpretation would be opposed to an interpretation of Dennett's story about the way intentional systems form a self-concept used in the prediction and explanation of their behaviour. This way of interpreting Dennett depends on emphasizing the fact that for Dennett an intentional system's self-concept is an abstract entity consisting of external attributions and self-attributions made as a response to the observed behaviour that this system displays. This kind of self-concept formed as a response to the intentional system's actions is not influencing them but it is only used in explaining them and predicting what it will do next.

What Dennett's account is missing, in Velleman's view, is an explanation of how a change in the self-conception of an individual can induce radical changes in the individual's actions. Both authors consider extreme pathological cases in which individuals act under more than one self-conception (cases of Multiple Personality Disorder or, as it is now known as, Dissociative Identity Disorder) and each of them draws different conclusions from them[62]. Dennett describes these cases as being occasions where the afflicted individual's behaviour is so complex that more than one self-conception had to be postulated in order to explain it or predict what the individual will do next. An interpretation of such behaviour relying only on one self-concept would be self-contradicting, while interpretation based on a number of different self-concepts would manage to be consistent. Velleman considers this view and agrees that more than one self-concept has to be postulated to understand this kind of behaviour. What he disagrees with is a view in which these self-concepts are

[62] See Nicholas Humphrey and Daniel C. Dennett, 1989, "Speaking for Ourselves: An Assessment of Multiple Personality Disorder.", *Raritan* 9(1), pp. 68-98, as well as Velleman, "The Self as Narrator."

only postulated as a way of understanding the individual's behaviour without influencing it in any way. His argument is that such a view doesn't provide any explanation for the reason these self-conceptions reflect every aspect of the behaviour of the individual under question. This correspondence can only be explained, in Velleman's view, only by assuming that these different self-conceptions actually influence this behaviour in different ways.

It might seem that this argument is pretty insubstantial. A reply could be that, yes, the different self-concepts do reflect all aspects of the observed individual's behaviour. That's the point. This is why these self-conceptions have been postulated in the first place, to explain the fact that the individual seems to behave in ways that cancel each other out. First, let's say, it might express a certain set of goals, beliefs, hopes, desires and other such states in its behaviour that it shortly thereafter completely undermines by expressing an incompatible set of goals, beliefs etc. Someone interpreting the behaviour in question achieves a consistent interpretation by postulating two different self-concepts expressing two incompatible sets of intentional states. The self-concepts do not have to influence the observed individual's actions, as long as they provide a consistent explanation of them.

My view is that this reply ignores the main worry that I believe lies behind Velleman's observation. What I take to be Velleman's main point when arguing that the incompatible self-conceptions reflect all aspects of the individual's behaviour is that there is a certain kind of unity in both sets of behaviour that can only be made sense of if the self-concepts expressing different sets of intentional states do influence these different incompatible sets of behaviour. As he writes:

> "Why should discontinuities in the patient's autobiography be accompanied by corresponding changes in the patient's course and manner of action? If a human being just contains "lots of subsystems doing their own thing", then why can't one of them do its thing with his feet just as another does its thing with his mouth, so that he walks the walk of one personality while telling the story of the other?" (Velleman, "The Self as Narrator", p. 9)

This just brings us back to our initial question. How can such unity be displayed by creatures that are made up by a complex distributed array of information-processing mechanisms? Dennett and Ismael seem to have an answer to this question, but the problem as we have seen is that it is not quite clear how the information contained within a self-concept (or more accurately in the narrative stream including the self-concept that models the agent's features and his location, *a la* Ismael) is used by the intentional system in its actions. I also think that the way Dennett and Velleman have been opposed in the preceding discussion is too simplistic, as it seems that Dennett also sees the self-concept as an important motivating factor on an intentional system's actions. But Velleman's arguments against Dennett do point towards some internal inconsistencies in Dennett's work, as he sometimes seems to claim that the unity displayed in behaviour does not depend on the self-concepts we postulate to explain it.

Velleman also tries to resolve this inconsistency, so I think that his account of narrative control can shed some light on the problem, assuming we recognize that human beings are, indeed, made up out of lots of subsystems doing their own thing. So, special care should be taken when postulating a "narrative module", as Velleman does, that leads to coherent narratives being formed by the agent that express his self-understanding by corresponding to his self-concept. A temptation that could create significant misunderstandings is to claim that the narrative module itself fits the agent's self-attributions to his actions according to his circumstances. Velleman himself seems to succumb to this temptation while describing the functions of the narrative module (especially in "The Self as Narrator"). But I think that the proper way of describing the module is not as an executive entity forming a coherent narrative for the agent, since that is what we have been trying to avoid all along. Instead, I think we should claim the agent's desire for self-consistency motivates him to form a coherent narrative by fitting his self-attributions to his actions, according to the circumstances he finds himself in. In this sense, we might say that the agent acts as an active narrator when expressing his self-concept in his actions. What motivates this agent to act as such is his desire to understand his actions as expressing his reasons for acting as such.

Following Velleman, the acting individual would not be able to act as an agent if it didn't have a drive to express its motives coherently in its actions. But this drive depends on the acting individual's having the capability to form such narratives from its actions, since they are not formed by the drive itself. Going back to Dennett and Ismael's terminology, the Joycean Monologue is formed by the Joycean Machine and the processes making up the Joycean Machine use the information contained in the Monologue in order to form a narrative that coheres with that kind of information. The Joycean Machine acts as a narrative module when it leads to such narratives being formed because these narratives express the agent's reasons for acting and the agent achieves a kind of self-understanding based on those narratives. But why do the narratives formed by the agent in action express his reasons for acting? And how is it exactly that these narratives are formed? We need to take a closer look to what the proposed account of narrative control entails.

For Velleman, the agent is a little like an actor improvising his role in a play.[63] There's no set script guiding his every move, no ready-made role that he must enact. The behaviour he exhibits in the play, such as the lines he delivers, is being made-up by him on the spot. It also takes the form of a response to the actor's ever changing circumstances within the play he participates in. Other actors also improvise their lines and movements and the actor has to take those into account when creating the role that he enacts. He tells stories about himself and his circumstances and he acts as the protagonist of these narratives. In the course of creating those narratives, the actor represents himself as having certain intentional states that play a causal role on his behaviour and are caused by various changes in his circumstances. In a sense, the actor's responses to his environment express his motives, which the actor attempts to express in his actions. When the actor expresses the motives he endorses in his behaviour, he knows what he's doing because he's doing what the persona in the role he enacts would do and knows what his motives are because they are expressed in his actions. His behaviour makes sense to him as behaviour that arises from the stories he tells about himself and his circumstances.

The next step taken by Velleman is to remove the line between the actor and the persona that he's playing. The agent is playing himself according to the narratives

[63] See e.g. Velleman, 2009, p. 14:
"Imagine an actor who plays himself, responding to his actual circumstances and manifesting the occurrent thoughts and feelings that the circumstances actually arouse in him, given his actual attitudes and traits.

This actor improvises just as he did when portraying a fictional character, by enacting his idea of how it would be understandable for his character to manifest his thoughts and feelings under the circumstances. But now the character is himself, and so what would be understandable coming from the character, given the character's motives, is what would be understandable coming from him, given the motives he actually has. Thus, he manifests his actual thoughts and feelings, as elicited from his actual makeup by his actual circumstances, in accordance with his idea of what it makes sense for him to do in light of them."

created in response to his circumstances. His self-attributions are made in response to his changing circumstances and they feed back into his behaviour as a guiding influence. Where does the narrative module come in? In my understanding of Velleman, the narrative module is constituted by the agent's desire to make sense of his actions and it integrates the agent's self-concept with his actions and his circumstances. But note that in this account, contrary perhaps to Velleman's understanding of narrative control, the desire for self-consistency does not act as a narrative module itself but it enables the agent to act as a narrator.

The narratives incorporating the agent's self-attributions influence his behaviour because the agent wants to achieve an understanding of himself through his behaviour. Returning to the actor analogy, if the persona the actor was playing was facing its arch enemy then the behaviour that would make sense as expressing this persona would be to treat the enemy as such, perhaps by attacking him. If the actor went on with the play by treating his enemy as his best friend then this kind of behaviour would presumably make no sense either to the actor or to his enemy. Excluding special explanations of what the actor is up to, the spectators of the actor's performance would agree that he has no idea what he's doing or what his motives are. In the same way, an acting individual acts as an agent by behaving in ways that make sense because they arise from the motives he endorses. His actions are explained as being the causal products of these motives. Going one step further, in accordance to Velleman's account, the actions the agent takes express his reasons for acting in such ways. That is because the agent's internal states (beliefs, desires, intentional states in general) express the ways he responds to his circumstances. The ways he responds to his environment, in turn, constitute his reasons for acting in certain ways. In terms of narrative, the agent creates narratives of his actions according to what his circumstances dictate and these narratives include his self-attributions of intentional states[64] and his representation of his environment. He subsequently enacts these narratives by behaving in accordance to them (if there is no weakness of the will or he doesn't try to enact narratives that do not express his reasons for acting).

[64] Note that these self-attributions are not the agent's intentional states themselves. They should more strictly be taken as judgments expressing the agent's intentional states.

Combining the account of narrative control inspired by Velleman's work with the account of self-regulation inspired by Dennett and Ismael's perspectives makes, I believe, for a better-rounded story of what it means for a human being to act as an agent. We are complex intentional systems consisting of a self-organizing substructure, with distributed processes carrying various kind of content with a more or less prominent role in our behaviour. We also model ourselves and our environment. This model takes the form of a self-concept consisting of information about our mental and physical features, which lies at the centre of a narrative stream consisting of information representing our environment. The narrative stream built around a representation of ourselves is constantly adapting to our changing circumstances, reflecting our developing responses to our environment. This means that our self-concepts are not fixed throughout our lives but change according to changes in our environment. This also has as a consequence that the way we respond to our environment changes along with the way we view ourselves and our circumstances. So it is more accurate to talk of different narratives formed around an evolving self-concept, depending on the ways the world we interact with changes.

The reason we regulate ourselves in accordance to our self-concepts might appear to be mysterious at first. A way to dispel this mystery is to reduce our agency to the function of a desire for self-consistency, which enables us to act as self-improvising narrators. We constantly narrate our actions and we try to keep our actions consistent with the narratives we form. That's because these narratives express our intentional states, which can be viewed as our dispositions to act in certain ways, while these states in turn express our reasons for acting because they express our understanding of what kind of responses are entailed by our circumstances. Acting in accordance with these motivating states enables us to achieve a kind of folk-psychological understanding of ourselves, because it enables us to understand what we're doing as explained by these motivating states, which are caused in certain ways and have particular effects in our behaviour.

Empirical studies

Several empirical studies seem to lend support to the claim that people tend to act according to the way they view themselves and their circumstances. Subjects of certain experimental settings have been manipulated to exhibit behaviour that corresponds to the feelings and intentional states most appropriate to these settings. Velleman himself has explored such studies in detail and believes that they offer indirect evidence in favour of his account[65]. The account elaborated thus far differs form Velleman's in some respects (most notably in avoiding the claim that the drive towards self-consistency can be viewed as a narrative module that itself performs all the functions of agency), but it also depends on the existence of the motive to regulate one's self according to one's self-conception, or the narrative drive towards self-consistency. It would be worth then considering whether Velleman is right in claiming that the existence of such a drive has been hinted at in relevant research in cognitive and developmental psychology.

We can separate the research examined by Velleman into three main areas. First, there is research into what has been called "cognitive dissonance", in which subjects are manipulated in order to engage in behaviour that does not seem to cohere with their motives at the time. The second is developmental research on whether children internalize specific characterizations of themselves that they subsequently express in their behaviour. The final strand of research involves the self-attribution of feelings and intentional states by subjects and the extent to which these subjects are influenced by these self-attributions. The main theme in the studies in all three areas seems to be that the subjects' self-concept can be influenced by the circumstances in the experimental settings. Whether this self-concept is in turn used by these subjects to influence their actions remains to be determined.

The phenomenon called cognitive dissociation describes cases in which subjects seem to change their mind about what they thought about something (such as an activity they engage in) because of a seemingly irrelevant experimental manipulation. For example, one such study, conducted by Festinger and Carlsmith (1959), involved psychology students in performing a task that was specifically designed to be bland and tiresome[66]. Some of these subjects were then offered the choice to report this task as being enjoyable to an associate of the experimenters that was pretending to be a subject waiting to perform the same activity. One of the

[65] See Velleman, 2000, "From Self-Psychology to Moral Philosophy", *Philosophical Perspectives* 14, pp. 349-377

[66] See L. Festinger and J.M Carlsmith, 1959, "Cognitive Consequences of Forced Compliance", *Journal of Abnormal and Social Psychology* 58, pp. 203-211.

groups was given a very low monetary award (one dollar) as an incentive to report that the task was enjoyable, while another was given a more substantial amount (twenty dollars). The subjects who were not paid to describe this activity as enjoyable formed the experiment's control group.

After performing this task (and for the low and high monetary award subjects, after reporting this task as being enjoyable to the associate of the experimenters), all subjects had to complete questionnaires in which they indicated the extent to which they enjoyed the task. As expected from the design of the initial task, the control group's ratings indicated that they did not enjoy performing the initial task. However while the high monetary award subjects' ratings were only marginally different to those of the control group, the ratings of the low monetary award group indicated that these subjects found the same task significantly more enjoyable than the rest of the experiment's participants. What seemed to be the case is that the subjects that were paid the least amount of money ended up apparently believing their reports that the task they had just performed was enjoyable[67].

An interpretation of this result that Velleman considers is that the low monetary award subjects could not explain why they would lie about what they thought about the task, since the reward they were given for lying wasn't really substantial enough as an incentive to lie. An assumption that Velleman recognizes as underlying this interpretation is that these subjects are unaware of the overall influence the experimental setting has on them, that is, they are unaware that this is a setting specifically designed in order to exert pressure on them to lie. But if they are unaware of that influence and they find themselves reporting that their task was pleasant only because of a low monetary reward, these subjects seem to face a conflict between two different understandings of their situation, one in which the task they just performed is dull and another in which it was enjoyable. In the first understanding these subjects might consider, the task was dull and the reward offered was not a strong enough incentive to lie, so they ought to have reported that the task was not enjoyable. The alternative understanding that might be adopted by these

[67] See Festinger and Carlsmith, 1959, p. 208:

"In short, when a [subject] was induced, by offer of reward, to say something contrary to his private opinion, this private opinion tended to change so as to correspond more closely with what he had said. The greater the reward offered (beyond what was necessary to elicit the behavior) the smaller was the effect."

subjects is that they did find the task enjoyable, in which case the low monetary award was sufficient since they were going to report that the task was enjoyable anyway. These subjects might then resolve this conflict and give meaning to their reports by adopting the alternative understanding of their situations and by forming the belief that they actually perceived the task as enjoyable all along. This way of seeing things makes their behaviour less mysterious for them, because it makes for a coherent explanation of their behaviour.

Velleman considers two ways in which this conflict resolution in the low monetary award subjects can be further described. According to the first, the subjects already had a belief that the task was dull which did not cohere with their subsequent reports that it was bland. According to the second, the low monetary reward subjects only formed one belief, which they attributed to themselves as the belief that would make sense according to what they say to others about their task. The difference between these two interpretations is that according to the first, the subjects replace their previous belief that the task was bland with a new belief that the task was enjoyable all along[68], while in the second the subjects do not change their mind about what they used to believe but they make up their mind about what their behaviour indicates about whether the activity they engaged in was enjoyable or not. In any case, Velleman's insight is that these interpretations do not necessarily have to be opposed but they might both describe the workings of the subjects' motive for self-understanding, since in both cases they adopt a self-conception that coheres with their behaviour.

In general, the results of these experiments seem to describe a pretty common psychological phenomenon. When we spend great amounts of effort on something, we give this effort some kind of meaning, perhaps in the sense that it leads to a result we find rewarding or maybe because we enjoy the effort in itself. It seems unlikely that someone would willingly spend a lot of energy, time and effort on something that they didn't feel was meaningful in any way. We might sometimes even change our minds about what the effort meant to us or how successful the results are, in

[68] An implied pre-requisite that Velleman also recognizes as needed for this interpretation to work is that the people who changed their minds are unaware of the fact that they did that. Otherwise it looks like they would express two simultaneously contradicting beliefs about what they thought about the task in the first place, which doesn't seem likely.

order to avoid concluding that the effort was a waste of time and energy. But I agree with Velleman that regardless of whether we do change our minds about what our efforts signified in the first place or instead we make up our minds about what it signifies depending on our behaviour at the time, it still seems that in general we are trying to explain our actions as resulting from a coherent self-conception. This result, however, indicates only one direction of fit, from the world to the mind. What the account we've been developing assumes is that we not only make up our minds according to our circumstances but we also behave according to our self-conception. Does the relevant research hint towards that direction of fit?

The developmental evidence examined by Velleman suggests that children manifest certain traits in their behaviour more often when these traits have been attributed to them by the experimenters. In one relevant study, for example, pupils of an elementary school were divided in three groups, with one group being told that they had a tendency to be tidy (the attribution group), another group being told that they ought to be tidy (the persuasion group) and a third group (the control group) that were not told anything[69]. What transpired was that both the persuasion and the attribution group showed an increased tendency against littering, when compared to both previous tidiness and the behaviour of the control group. But the pupils in the attribution group, that were told they were tidy instead of urged that they ought to be tidy, showed the most prolonged tendency to act according to the attribution of tidiness. Further similar developmental evidence examined by Velleman suggests that younger children (around 5-6 years old) do not respond in the same way to such attributions, while older children do[70]. This result has led the experimenters to hypothesize that an ability to understand what having certain traits of character implies is required before the traits are suitably implemented in one's self-conception so that they are manifested in one's behaviour. An ability of this kind would presumably consist in the understanding that such traits tend to be stable over time and to manifest themselves in various specific ways.

[69] See R.L. Miller, P. Brickman and D. Bollen, 1975, "Attribution versus Persuasion as a Means for Modifying Behavior", *Journal of Personality and Social Psychology* 31, pp. 430-441.

[70] J.E. Grusec and E. Redler, 1980, "Attribution, Reinforcement and Altruism: A Developmental Analysis", *Developmental Psychology* 16, pp. 525-534.

Apart from such developmental research hinting at the extent to which self-attributions of character traits influence subjects' behaviour, Velleman also presents a wealth of studies whose results can be interpreted as indicating that subjects attribute certain motives to themselves (such as certain emotional states) that they subsequently enact in their actions[71]. In a representative example from these studies, Zillmann and his co-authors (1974) arranged for subjects to participate in an experiment whose first stage involved engaging in a learning exercise with a confederate whose hidden goal was to anger the participants. This initial interaction took place over an intercom and the subjects were led to believe that the purpose of their learning exercise was to teach the confederate to avoid certain errors by administering electric shocks of varying intensity in response to the confederate's mistakes. After engaging in this activity, the subjects were instructed to express their opinion on a list of topics to the confederate, who could either indicate his agreement by turning on a light signal or indicate his disagreement by administering shocks to the subjects. Regardless of the opinions expressed, the confederate always administered shocks 9 out of 12 times. This was hypothesized by the experimenters to be sufficient to invoke the subjects' anger towards the confederate.

Following this initial interaction, the participants were led to engage in rigorous exercise on a training bicycle while having their heart rate and blood pressure measured and a selection of slides shown to them, with the pretence that they were being tested for their memorization skills under conditions of physical stress. Some of these subjects exercised for 1.5 minutes followed by 6 minutes of sitting, while the rest performed this procedure in reverse order, concluding with the short burst of strenuous exercise. The final stage of the experiment had all subjects return to the learning exercise, in which the confederate was instructed to make a certain amount of errors to which they had to respond with the administration of shocks.

Zillmann and his co-authors' hypothesis was that the subjects who concluded the previous stage of the experiment with the short burst of exercise would deliver

[71] See e.g. S. Schachter and J.E. Singer, 1962, "Cognitive, Social and Physiological Determinants of Emotional State", *Psychological Review* 69, pp. 379-399, D. Zillman, R.C. Johnson and K.D. Day, 1974, "Attribution of Apparent Arousal and Proficiency of Recovery for Sympathetic Activation Affecting Excitation Transfer to Aggressive Behavior", *Journal of Experimental Social Psychology* 10, pp. 503-515 and D. Zillman, 1978, "Attribution and Misattribution of Excitatory Reactions", J.H Harvey, W. Ickes and R.F. Kidd (eds.) *New Directions in Attribution Research,* Vol. 2, pp. 335-368.

shocks of lower intensity to the confederate than the subjects who were given some time to rest, since the former group would be more likely to attribute their state of arousal to the immediately preceding exercise. In contrast the subjects who were given some time to rest were hypothesized to respond with shocks of higher intensity to the confederate's errors, as they would attribute their state of arousal to anger rather than to the effects of their exercise. The study's results lend themselves to a fit with this hypothesis. The subjects who were allowed to rest did respond with shocks of a higher intensity to the confederate's errors, both compared to their previous responses in the initial stage of the experiment and to the responses of the subjects who had just exercised. This behaviour seems to fit the interpretation according to which the subjects who had time to rest attributed their state of arousal to residual feelings of anger towards the previous provocation they received, which led them to manifest increased aggression towards the confederate as the object of their perceived anger.

In line with Velleman's general argument, these results indicate that we not only adjust the way we view ourselves because of changes in our circumstances, in order to maintain some kind of consistency between these circumstances and our self-conceptions, but that we also enact the way we view ourselves based on our circumstances in our actions. Based on such research then, the idea that we have a motive that leads us to try and form coherent narratives integrating our self-conception, our actions and our circumstances in a whole does not seem so outlandish. But even so there is a significant caveat (given considerable attention by Velleman) that might frustrate an interpretation that postulates a motive for self-consistency underlying our actions as agents.

The caveat is that there is a "self-enhancement" interpretation of these results that seems to cast doubt on Velleman's account. On the self-enhancement interpretation, people are not motivated by the need for self-consistency, as Velleman suggests; rather they are simply motivated by the need to maintain a positive self-conception[72]. In the experiments discussed above, the self-enhancement approach would explain the results as follows: The subjects in the cognitive dissonance studies don't want to

[72] See e.g. C.M Steele and T.J. Liu, 1983, "Dissonance Processes as Self-Affirmation", *Journal of Personality and Social Psychology* 45, pp.5-19.

look like they have no idea what it is they're doing, so they seek out a self-concept that doesn't contain any contradictions, while the children in the developmental studies seek to propagate a view of themselves according to which they have what is usually perceived as a positive character trait, that of being tidy. This interpretation though falls short when it comes to the self-attribution of motivating states such as anger, as Velleman argues. Anger is not usually viewed as a positive character trait but is instead an emotional state that does not necessarily have any positive or negative value associated with it. And yet the subjects in such experiments still seem to act according to the way they perceive themselves as feeling, or according to the intentional states that they perceive themselves as having. Hence, self-enhancement might be a viable interpretation of a part of what's going on when people act according to their self-conception but it doesn't seem to be the whole story.

Velleman also defends his view by referring to studies according to which people tend to propagate negative self-conceptions of themselves by going as far as to act in ways that will make others adopt a similar view of them as they have of themselves[73], but I think that some support can be found against the self-enhancement hypothesis even in the aforementioned examples in which it seems to get things right. In the cognitive dissonance example, the way Velleman characterizes the interpretation of the subjects' behaviour according to the self-enhancement hypothesis is this:

> "According to dissonance theory, the….subjects came to believe what they had said in order to escape a specifically cognitive predicament, of being unable to explain their behavior, or of finding it contrary to expectation. But they might instead have come to believe what they had said in order to escape the appearance of having been irrational, in having said it for no good reason. In that case, their change of mind would have aimed to rationalize their past behavior rather than to remedy their current state of reflective ignorance or incomprehension; and it would thus have aimed at removing not a cognitive problem but a threat to their self-esteem as rational agents. The effects of forced compliance have therefore been taken by other psychologists to indicate a motive for attaining a favorable view of oneself rather than for maintaining

[73] See e.g. W.B Swann Jr. and S.J. Read, 1981, "Self-Verification Processes: How We Sustain Our Self-Conceptions", *Journal of Experimental Social Psychology* 17, pp. 351-372, W.B Swann Jr. and C.A. Hill, 1982, "When Our Identities Are Mistaken: Reaffirming Self-Conceptions through Social Interaction", *Journal of Personality and Social Psychology* 43, pp. 59-66 and W.B. Swann Jr., C. De La Ronde and G. Hixon, 1992, "Embracing the Bitter "Truth": Negative Self-Concepts and Marital Commitment", *Psychological Science* 3, pp. 383-386.

> consistency with one's actual self-view—a motive of self-enhancement rather than self-consistency." (Velleman, 2000, p. 357)

In my view, avoiding a threat to the self-esteem of the subjects as rational agents does not necessarily support a motive towards self-enhancement. What this tendency indicates is that the motive for self-consistency operates within a framework in which creatures like us act as rational agents that know what they think and know what it is they're doing, while also being seen as such by other agents who can in turn anticipate and rationally respond to these actions. Inconsistency in our actions and intentional states is disruptive to our status as agents and as such, a motive for self-consistency is essential not only for us to understand our actions as coherent products of rational, responsible agency, but also for efficient social collaboration with other agents[74]. Simply taking our attempts to avoid behaving as if we have no knowledge of and control over our minds and actions to indicate a motive for self-enhancement obscures, in my view, the bigger picture wherein our agency in our minds and actions is expressed by maintaining consistency between what we think and what we do.

As for our example of developmental research, it seems that the children in the case study are not only motivated by a desire for self-enhancement. Again, a clue for why that is so might be found in Velleman comments of the results of this study:

> "These experiments compared favourable attributions with injunctions, which did not have a similarly favourable tone and might even have been interpreted by the children as presupposing an unfavourable attribution instead. (Why would teacher exhort us to be tidy if we weren't in fact untidy?) Perhaps, then, the experiments demonstrated, not an interesting motivational difference between attributions and injunctions, but an utterly unsurprising difference between negative and positive reinforcement." (Velleman, 2000, p. 359)

Remember that children told they were tidy went on acting as such more than children who were told that they should be tidy or children that weren't offered any suggestions. A proponent of self-enhancement might argue that the children

[74] See the discussion in previous chapters on the role played by our training in folk psychology in the development of our agency. As we've seen, only in this framework does a desire for self-consistency become essential for our capacity to act as agents. In an environment where we are trained to act in a manner that is intelligible both to ourselves and to one another, a motive to maintain consistency between our self-conception and our actions is psychologically effective because it enables us to appropriately respond to the social norms that give structure to our on-going interactions.

preferred a favourable conception of themselves as tidy, but as Velleman suggests, the children who were told they ought to be tidy might have interpreted that suggestion as an indication that they were untidy. So, according to this explanation, these children didn't maintain a tidy behaviour as much as the group of children who were told they were tidy because they had a self-conception of being untidy. Since this self-conception is not a favourable one and yet it still seems to have had an influence on the children's behaviour nonetheless, even in this case an interpretation according to which the children acted only under a motive for self-enhancement and not for self-consistency is found to be wanting.

In general, I think that studies such as the ones reviewed by Velleman can indeed provide some support for the hypothesis that we all have an integral motive for self-consistency, or self-understanding (and not just for self-enhancement) that enables us to fit our minds to our actions and our actions to our minds. This motive, as we have seen, is essential for our account of human agency, so even though the relevant psychological literature does not provide indisputable proof in favour of this account, it at least indicates that it might not be a complete philosopher's fiction either.

Language and self-regulation

Let's take a look at how the story goes so far. In attempting to explain human agency, I started with examining the behaviour of simple distributed information-processing systems such as the ant colony. Such systems can be called intentional since they appear to engage in purposeful behaviour that can be explained and anticipated by using Dennett's intentional stance. Their behaviour can be explained by reference to the goals they are pursuing and to the means through which they pursue these goals. One can even talk of the system's attitudes towards its environment and use these attitudes as theoretical constructs that help with anticipating what the system's next moves will be and explain why it behaved in those ways. The ant-colony depends on individual distributed information-processing mechanisms (the ants). Postulating certain goals and attitudes that this system expressed in its behaviour (for example, the goal to protect itself from its predators and its belief that there is a predator nearby) can help with explaining its behaviour and predicting what it will do next, depending on how its environment changes. This also has the effect of an appearance of purpose and unity being created in the observer of such behaviour.

The insight we got from self-organization is that in systems such as the ant colony there is no specific central control mechanism which contains the attitudes and goals that the user of the intentional stance postulates. Instead the ants organize themselves into a whole by responding to each other and to their surroundings based on their innate dispositions, which are the products of evolutionary conditioning. The obvious step here seems to be to extend the case of such systems to our case, arguing that human beings are also self-organizing, distributed intentional systems that differ from the simpler self-organizing systems because of their complexity. The means by which humans represent their environment and respond to it are a lot more varied and complex that the means of a system such as an ant colony. But this does not mean that there is a central executive mechanism in humans which is responsible for pursuing certain goals and choosing the means by which to pursue them, in accordance to their intentional states. Authors such as Dennett argue that the analogy with the ant colony and our continuum with simpler intentional systems should be taken seriously. We've also been conditioned to behave in certain ways by cultural and biological evolution and the behaviour we display is the product of the interaction of a variety of self-organizing sub-mechanisms.

The problem with this simple analogy is that it might obscure an important difference that arises from an increase in the variety of the means with which we represent and respond to our environment. I've argued that we should recognize that increased complexity in self-organizing intentional systems brings with it genuinely new and interesting capacities, while still not quite necessitating the existence of a central executive mechanism in the form of a Cartesian self that is responsible for the system's purposeful, unified behaviour. What I think is the main difference between simple self-organizing systems and more complex ones is the ability to navigate one's environment using a representation of one's self as a guiding influence. Ismael's work allowed us to see how complex, distributed intentional systems might be said to regulate their behaviour according to their self-conception. Self-regulation in this view consists of the system's being able to use some of the input it receives in order to form a unified informational stream, which Ismael calls the Joycean

Monologue, that contains an objective representation of its environment built around a representation of the system's main features that are relevant to responding to its surroundings. Such features might include the system's long-term goals and intentional states, as well as the system's physical capabilities. We've seen that such a system can display significant flexibility in the way it responds to its environment because of its evolving representation of itself.

While Ismael's story gives us an initial understanding of the genuine differences between creatures like us and simpler self-organizing systems, I've argued that it should still not be straightforwardly applied to the case of human agency. One reason for that is that there is a vagueness in the story as told so far that makes it mysterious how agentive processes such as long-term planning and deliberation can be performed by self-regulating creatures. At times, the danger of accounts such as Ismael's is that they seem to give the unified information used in self-regulation too much authority, so that it sounds like the self-concept itself, or the virtual map containing that information, exercises some active influence on the system's actions. What agency in one's action amounts to is not easy to identify in such an account. I've argued that we should see the entire system as expressing its agency in its actions and not that there is a part within the system that acts as an agent and expresses its active contribution in the system's actions.

Velleman's account allowed us to see how we can provide a reductive account of agency that does not posit the existence of an all too knowing component in the self-regulating intentional system. The idea is that an agent acts as such because of his motive for self-consistency, which allows him to form certain narratives of his actions as corresponding to his self-concept and ensure that there is a fit between these narratives and his actions. With the help of such an account it becomes clearer how an agent forms a self-concept that he uses in order to express his active contribution, or first-person perspective, in his actions.

The problem is that the jump from self-regulation to human self-regulating agency still seems strenuous. After all, a simple answer to how and why self-regulating systems use their self-representation to guide their actions is that they're just built, or wired, that way. Why does Ismael's self-regulating ship use the information contained in the map it uses in order to determine its next destination and follow the course that will lead it to it? That's just the way it's been designed. Besides, another example of a self-regulating system Ismael uses is that of a missile tracking its target using information contained in its tracking system, which also contains information representing the missile's features that are relevant in the course it follows. There's nothing particularly mysterious about how such a system functions. Why can't we extend these analogies all the way and say that we act as self-regulating systems because we've evolved to act as such? Why not extend that to an account of human agency and say that agents act as such because of the way they've evolved?

One answer is that some of the phenomena we take to be associated with human agency, such as agents holding each other responsible for their actions, would still be mysterious under a direct analogy of self-regulating systems with human agents. The defenders of such an analogy might either say that responsibility is an issue that is outside the scope of an account of self-regulation and that some of the features we regularly associate with human agency could turn out to be social constructs that have no bearing on the way humans are empirically discovered to behave. Regardless of how one feels about issues of responsibility and what role they play on accounts of human agency based on self-regulation, a problem that I think is more pertinent to a direct analogy between self-regulating systems in Ismael's sense and human agents in our sense is that it would miss the profound difference the use of a common language that intentional systems use (which can be used to convey their common understanding of what it means to express one's agency in one's actions) makes for the way these systems behave. More specifically, I think that the use of a public language in which a folk-psychological understanding of agency is couched makes for a genuine difference between intentional systems such as self-regulating ships and human beings.

Two authors that I think appreciate this kind of difference are Victoria McGeer and Philip Pettit. In their article "The Self-Regulating Mind", McGeer and Pettit argue that we should distinguish what they call "routinised" intentional systems from self-regulating ones and that the human mind belongs to the latter category[75]. Routinised systems are basically those systems to which the intentional stance can be fruitfully

[75] See Victoria Mc Geer and Philip Pettit, 2002, "The Self-Regulating Mind", *Language and Communication,* Vol.22, no.3, pp.281-299.

applied. These systems can be viewed as having certain intentional states that they express in their behaviour and as making use of various means in order to achieve their goals. While acknowledging the influence of Dennett's intentional stance, these authors describe a routinised intentional system as such:

> "To be an intentional system, and therefore qualify as 'minded' in some minimal sense, is, on standard approaches, to be a system that is well-behaved in representational and related respects. The well-behaved system represents things as they appear within the constraints of its perceptual and cognitive organisation. And it acts in ways that further its desires—presumptively, desires that reflect its overall needs and purposes—in the light of those representations or beliefs." (McGeer and Pettit, 2002, p.282)

According to this way of ascribing intentionality, humans, ant colonies and self-regulating ships are intentional systems. They all can be viewed as having intentional states and goals that they express in their behaviour. The difference between humans and other kinds of intentional systems, according to Pettit and McGeer, is that humans can express their intentional states and goals in a shared language. This gives them the possibility to regulate their behaviour in a way that is unavailable to other intentional systems, because by expressing states such as beliefs, they can also recognize the content of such states and also recognize what having such states commits them to. In other words, the argument that these authors use to distinguish humans from other intentional systems is that humans are intentional systems that can recognize the fact they are such systems and include that fact in the way they represent the world. And because such systems can recognize the states that play a role in their behaviour, they can also recognize the relevant constraints that behaving in such ways commits them to.

What is an example of such pressures under which intentional systems behave? Focusing on the case of belief, McGeer and Pettit argue that "the intentionally well-behaved system must tend to believe the true; must tend to believe only consistent contents; must tend to believe contents that are inductively or deductively supported; and so on" (McGeer and Pettit, 2002, p.287). Identifying such constraints makes possible responses to them available. In these authors' view creatures like us, being self-regulating systems of this kind, can make sure that we conform to such constraints of rationality by treating our beliefs and the evidence for our beliefs in various ways, such as making sure that we do not retain a false belief even though it might sometimes be hard to do so.[76] Even though they mainly focus on the case of belief, I think that these authors intend their discussion to apply to all the mental states that can be expressed in a shared linguistic framework.

Another consequence of using a shared linguistic framework to represent one's intentional states and a common folk-psychological understanding of intentionality couched in this framework, as McGeer and Pettit stress[77], is that intentional systems can learn to anticipate and respond to each other as intentional systems. A common folk-psychological framework that expresses an understanding of how a "well-behaved", or rational, intentional system must act in order to be viewed as such makes it possible for users of such a framework to hold each other and themselves responsible for such actions and to expect behaviour that is compatible with the constraints identified by this kind of understanding. For example, using McGeer and Pettit's example of rational constraints on belief, if we tend to believe what's true and well-supported by the available evidence and not to hold contradicting beliefs, then we will expect each other and ourselves to have such beliefs in order to be rational and to be able to respond to each other as rational believers. Hence, language and folk-psychology critically influence the way we view ourselves and each other and the way we respond to our environment[78].

From self-regulation to human agency

My view is that by using the aforementioned insights we are in a position to understand why self-regulation in Ismael's sense is not enough for a full account of human agency and why Velleman's work can help with making such an account

[76] Examples mentioned by the authors are attempts to avoid the gambler's fallacy and the effort that pilots in training make in learning to trust the instruments on their cockpit rather than their gut feelings. See McGeer and Pettit, ibid*,* pp. 289-290.

[77] See especially ibid, pp. 294-297.

[78]In his 2003, *Being No One: The Self Model Theory of Subjectivity,* MIT, Thomas Metzinger also explores the significance of our social nature for our responses to each other and our environment and makes the following claim:

"After language was available, we could not only communicate about…new facts and concepts but also proceed to publicly self-ascribe them to us...We started to consciously experience ourselves as thinkers of thoughts and as speakers of sentences…We started to *think* about ourselves as thinkers of thoughts and speakers of sentences…We experienced ourselves as individual beings that, at least to a certain degree, also were *rational subjects*…we could now also begin to make the fact that we are *social subjects* globally available for attention, cognitive processing, and action control...We were able to mutually *acknowledge each other as persons,* and to consciously experience this fact…Persons are never something we find *out there*, as parts of an objective order. Persons are constituted in societies. If conscious self-modelling systems *acknowledge* each other as persons, then they are persons." (Metzinger, 2003, pp. 599-601).

more robust. Ismael's ship is self-regulating in the sense that it uses information relevant to its situation centered on a representation of itself to navigate its environment. But it is not self-regulating in McGeer and Pettit's sense because it does not have the ability to express its states in a linguistic framework which would enable it to recognize what having such states amounts to. Such a self-regulating ship can use the information contained in the Joycean Monologue because it has been designed that way but it cannot recognize the fact that it's self-regulating in that way and so cannot respond to this fact. In that sense, the ship is closer to McGeer and Pettit's routinised systems since it operates under certain constraints that it cannot recognize and so cannot respond to in any way. The way it expresses its agency in its actions is different from the way human agents can, since such agents have a common understanding of what it means to express one's agency on one's actions and can regulate themselves in ways compatible with this kind of understanding.

Human agency might be different in important ways from other kinds of self-regulation, but I should make clear that it also depends on the kinds of self-regulation exhibited by systems such as the crewless ship. What both kinds of self-regulation share is the ability to navigate one's environment using a self-model. Both kinds of self-regulation also depend on a self-organizing substructure that is essential to forming a model of one's self as interacting with the environment. The crucial difference between the two is in how the self-ascriptions making up the system's self-concept are expressed. In the case of human agents, such self-ascriptions can be expressed in folk-psychological terms within a public medium. As we've seen in previous chapters as well, folk-psychology engenders a specific understanding of what it means to act as an agent and of the implications such actions have on other agents who need to anticipate and respond to them.

Considering these differences, there are several reasons that I think a motive for self-consistency and a story according to which human agents act as self-improvising actors is useful. One reason is that reducing the functional role of agency to a specific motive for self-understanding is a solution that avoids the problem of having to postulate one single element in the intentional system that is responsible for its expressions of agency. As I've argued above, we want to give an account of human agency that respects the fact that human beings are made up out of a distributed self-organizing substructure and such an account is not very hospitable to the idea of a single element in the organism having too much of a say on how it acts. Another reason is that the agent’s motive to fit the way he sees himself with his actions according to his circumstances fits with the view that human agents share a common understanding of what it means to be a “well-behaved” intentional system that enables them to identify and respond to the various constraints associated with acting as such. Such rational pressures are identified in the agent’s self-concept, since the agent’s self-ascriptions can be expressed in folk-psychological “mentalistic” language that enables the agent to recognize himself as being subject to certain demands of coherency.

In accordance to the account we’ve been developing, in order to understand his actions as expressing his agency, the agent has to act in a way that can be explained by what he takes to be the most appropriate response to his circumstances. What he takes to be the most appropriate response to his circumstances is reflected in his self-concept and is shaped by his identification of the constraints that he has to adhere to in order to act as an individual that knows its own mind and knows what it is it’s doing. Such an understanding is uniquely shaped by the use of a common folk-psychological framework couched in a medium that is shared by human agents. The use of such a medium and the common expectations it gives birth to is what sets human agents apart from other self-regulating intentional systems.

Chapter 4
The Limits of Conscious Awareness and Control

Introduction

As I've made clear in the preceding chapters, I favour an account of human agency according to which we are self-regulated creatures that are able to engage in actions expressing the attitudes we hold because we acquire a folk-psychological understanding of ourselves and we regulate our behaviour according to it. Human agency is distinctive because it has developed within a social framework wherein such interacting self-regulating creatures have the ability to recognize that they are self-regulating and respond to that fact in various ways, while also expecting others to have that capacity. These common expectations lead to certain normative standards that agents have to adhere to in order to maintain their status as rational, self-controlled persons that know what it is they think and do. Agents have to be able to explain their actions to themselves and others in ways that are acceptable, depending on the normative standards that have evolved through their interaction. In other words agents can provide reasons for acting in the ways they do.

Because agents can provide reasons for their behaviour, they can be held responsible for what they think (their attitudes toward the world that are expressed in their intentional states, emotional responses and other attitudes) and what they do (their actions which express their various attitudes). Furthermore, these self-regulated individuals are motivated to fit the way they see themselves to their actions because of their desire to understand themselves as acting in ways that can be explained by reference to certain reasons. The way these agents see themselves is crucial because it expresses their attitudes, which in turn can make actions expressing these attitudes intelligible by reference to these attitudes. Such a desire to understand oneself as acting for reasons has evolved because of the interaction among self-regulated creatures that regularly expect each other to provide coherent explanations for their behaviour. This view is combining elements from various authors' work[79] and a more robust support of these main elements can be found in the previous chapters. The key idea that I want to hold onto for this chapter is that of human self-regulation and the capacity for self-control manifested by fitting one's self-conception to one's actions.

My intention in this chapter is to test this account of human agency against the view that our behaviour is shaped through the influence of a variety of factors which frequently elude our conscious awareness. This influence consists in the activation and operation of unconscious processes that affect our perception of our environment and the way we respond to our circumstances. This line of thinking is becoming increasingly more popular in cognitive science and related fields of study, due to a variety of interesting findings in experiments testing the extent to which unconscious factors play a role in our behaviour. For example, findings in studies conducted by John Bargh and his co-workers suggest that there are certain processes, such as the pursuit of certain goals in action, which can be activated without a conscious effort on our part to activate them[80]. In the relevant literature, behaviour that is the result of conscious, effortful control is juxtaposed to unconscious, automatic response to one's environment[81]. There is considerable support for the claim that we do exhibit a kind of behaviour that consists in such unconscious responses, which are based on the activation of features such as unconscious goals and stereotypes influencing perception and action.[82]

The most exciting and potentially worrying implication of these studies is that this kind of behaviour is not as rare as we might initially suppose but is in fact highly pervasive throughout our every-day lives. We are constituted by distributed processes that can frequently guide our actions without the need for any conscious effort on our part. As an illustrative example we'll encounter further on when

[79] Especially McGeer 1996, 2001, 2007a, 2007b, Moran 1997, 1999-2000, 2001 and Velleman 1992, 2000, 2009 and"The Self as Narrator".
[80] See e.g. Bargh et al., 2001 and Bargh 2005.
[81] See especially Bargh and Chartrand, 1999a.
[82] See e.g. Bargh, 2005 and 2006, Bargh et al., 2001 and Bargh and Chartrand 1999a and 1999b.

examining empirical studies in more detail, our perception of certain groups of people can be influenced by common stereotypical features that they share, without us necessarily being able to consciously recognize that the way we perceive these people is shaped by our responses to these features. At the same time, the way we behave around people with these features is guided by the way we perceive them. It seems then that we might often respond to a group of people in similar ways without being able to consciously identify the factors that lead to our responses[83]. Also, as research conducted by authors such as Daniel Wegner indicates, we might also be mistaken when we try to trace our actions to their causes and have the feeling that our conscious intentions have caused an action that was in fact caused by unconscious processes that we are unaware of.[84] These processes might not only shape the way we respond to our environment but also the conscious thoughts that we have about these responses, which instead of being the cause of our actions might just be an epiphenomenal by-product of our behaviour.

This strand of research seems to suggest that the proponents of an account of agency according to which the way we see ourselves is instrumental to the self-control we exhibit in our actions would have to recognize that the way we see ourselves is shaped by factors that might lie beyond our conscious awareness. At worst, an implication we can draw from the aforementioned empirical studies is that we are driven, somehow, by the various forces in our world and that we are not really agents because we are not really in control of our actions. I think such a conclusion is overly simplistic because it is based on a flawed interpretation of the relevant evidence. First, this conclusion would overestimate the influence unconscious processes have on our behaviour and over-simplify the nature of these processes, while not giving enough weight to a view according to which both conscious reasoning and unconscious processes are part of a complex interplay guiding the ways we respond to our environment. Second, it would underestimate the extent to which human agency is a complex, holistic phenomenon that emerges from the social framework in which human agents respond to their environment and to each other. The account of agency I've been developing can be used in support of the claim that we exhibit a distinctive kind of self-control in our actions and that we can be held

[83] See e.g., Devine, 1989.
[84] See Wegner and Wheatley, 1999 and Wegner, 2002.

responsible for them, while also respecting the growing empirical evidence hinting at the limits of our conscious awareness of the causes of our actions and at the widespread influence our surroundings have on the way we respond to them.

More specifically, there are two morals that I want to draw in this chapter. First, we shouldn't ignore the evidence supporting the widespread influence of unconscious factors on the way agents think or act and on the way they see themselves and express themselves in their actions. We should also recognize that the way agents think and act and the way they see themselves and express themselves in their actions is dependent on the social framework in which their understanding of agency develops. Human agents learn to act as such because of their shared social practices that depend on communication and a shared public language. In order to engage in these social practices which lead to the development of the folk-psychological self-understanding and self-control that is unique to participants in these social "games", human agents need to use a common framework. Agents interact in a common framework through the conscious use of a public language, for how can communicating through language not depend on the agent's explicit knowledge of how language is used? The way agents understand themselves then is conscious in that it also depends on their explicit knowledge of how to communicate with each other using a common language and their explicit understanding of the information transmitted through such communication. This understanding leads to a conscious self-concept that expresses the agent's judgments on what his circumstances entail. This self-concept is also influenced by a variety of unconscious factors but the role these factors play in the agent's self-understanding and actions is in turn influenced by this self-understanding.

In order to further elaborate this account, I'll start by examining some of the empirical research aiming to enrich our understanding of the way we unconsciously respond to our environment and by considering possible implications of these studies. A prominent role will be given to Bargh and his co-authors' studies on processes such as unconscious goal activation and pursuit[85]. I will then examine Philip Pettit's attempt to place these findings in the perspective of an account of

[85] See especially Bargh et al., 2001.

human agency as a holistic, emergent social phenomenon[86]. While I agree in general with this way of seeing human agency, I'll argue that such an account needs to be elaborated with care so that it doesn't end up simply ignoring the available evidence because they don't fit a view according to which we have the capacity to exercise our agency in our actions. To this end, I'll extend our discussion by delving into more examples drawn from the surrounding interdisciplinary research that can support Pettit's view. Support of this kind can be found in the work of authors such as Daniel Dennett[87] and J. David Velleman[88], who each insists in his own way on the importance, for the development of human agency, of the capacity to give reasons for one's actions. Finally, John Haidt's theory on the interplay between conscious and unconscious factors in the production of moral judgment and Bargh's latter view on the interplay between conscious and unconscious processing will also be presented as example of views in which conscious factors can still influence our capacity to express our agency in our actions, despite the extent to which our actions are shaped by unconscious processes of which we are not consciously aware and we do not consciously control[89].

Questionable answers: setting the stage

The key features of what I take to be the best explanation for the way we express our agency in our actions should be restated. I think that our ascriptions of responsibility and intentionality to one another are not fundamentally misguided, since we do have the capacity to express our various intentional states in our behaviour in an authoritative manner. We do that by forming judgments on what our circumstances dictate, which function as normative commitments motivating us to manifest the intentional states expressed in these judgments. What gives these judgments their motivating force is our need to understand our behaviour as the product of our own authoritative agency. The way to do that is to understand our behaviour as expressing the reasons we have for acting in certain ways and these reasons are expressed in the judgments we make on what our circumstances entail. If

[86] See Pettit, 2007.

[87] See Dennett, 1993, 1996, and Dennett, 2003, *Freedom Evolves,* Penguin Books.

[88] See Velleman, 2009.

[89] See Bargh, 2005 and Haidt, 2001.

our behaviour reflects these reasons then it reflects our judgments which constitute our self-concepts.

The self-concepts we form are also dependent on folk-psychological attributions of intentional states, since we have been trained to use a folk-psychological framework in order to ascribe intentionality to ourselves and others. From a young age, we have learned to use explanations that refer to intentional states as propositional attitudes and to accept the implications of having such attitudes. For example, we have an implicit understanding of how certain beliefs interact with certain desires (e.g. the belief that a harpsichord is in the next room in combination with the desire to play the harpsichord can lead one to go into the next room). We also usually avoid knowingly holding contradictory attitudes and we tend to make an effort to accept the attitudes that we are aware of as following from the attitudes we already hold, so that we don't appear irrational. This kind of understanding informs our self-concepts, so we tend to understand the attitudes we hold as subject to these kinds of norms.

The attitudes we hold are behavioural dispositions which we manifest in a variety of ways, in our words, thoughts and actions. Such dispositions are not always automatically manifested in our behaviour and they require an effort to maintain. This does not mean that we just adopt whatever attitudes we choose. The attitudes that we can actively manifest in our behaviour are limited by what we judge as being the case. We try to maintain these dispositions because if we didn't, our behaviour would be mysterious in the sense that it wouldn't reflect our judgments on what our circumstances entail. It makes little sense to claim that one can have an attitude that one never expresses in any way in their behaviour. How can someone genuinely believe that what they take to be a wax apple is a real apple and act accordingly by taking a bite out of the wax replica they're holding? How can someone genuinely want to take a swim when they hate swimming and they'd never go anywhere near water if their life depended on it? And can someone intend to fly for Paris the next day when they take it to be the case that Paris has been obliterated in a nuclear attack? Someone can act as if it was the case that a wax replica is an apple, but if they know that it is made of wax it is hard to see what it would mean to say that they genuinely believe that it's real and they don't just act as such for a different reason (perhaps they're part of a play where they enact the action of eating an apple).

What I'm interested in is the case in which the agent actively expresses such attitudes in his behaviour in a self-knowing, self-controlled manner. I take that to imply that the agent knows what he's doing and why. The agent can take a bite out of a wax apple thinking it's an apple, but I find it hard to understand how an agent might still believe that the apple he's holding is real if he takes it to be made out of wax. If he keeps eating it and insists on it being a real apple, I think it would be plausible to say that he acts despite himself, which in this case would mean despite his best judgment of how the world is. Perhaps he's driven by a compulsive habit of finishing everything he's started, or he doesn't know what an apple is and what it means to eat a real apple if he tastes the wax but still insists that he's eating a real apple.

Acting in ways that are consistently opposed to the ways we judge we should act makes for behaviour that is confusing, both for us and for other observers attempting to interpret this behaviour based on our judgments. As in the wax apple case, the interpretation of a behavioural display as that of an agent who can be said to know his mind and control his actions might, if systematically problematic, be discarded in favour of a different kind of interpretation that attributes the display in question to causes external to one's agency. Because of the danger of losing our status as agents, which would imply that our claims about our actions would be ignored and that we're not in control of our actions, we have the tendency to act in a way that conforms to our judgments about how we should act. Since we've been trained to live up to these judgments by manifesting the corresponding intentional states in ways understood from the perspective of folk-psychology, we recognize that we should act in ways consistent with this framework in order to act as agents. Our judgments constituting our self-concept function as normative commitments that are used in regulating our behaviour in ways consistent with these commitments.

The main idea is that self-regulation is regulation in accordance to one's self-concept. This self-concept expresses one's judgments that constitute the reasons one has for acting, reasons that can be used in providing intelligible interpretations of one's behaviour. Interpretations that are intelligible are preferred by us because they facilitate our interactions with other agents and allow us to understand our actions as our own, in the sense that these actions are the enactment of our reason-responsive attitudes. The need to make our actions intelligible has been ingrained in our way of life to such an extent that it motivates us to develop the attitudes expressed in our self-concepts. Since we not only give these intelligible interpretations to each others' behaviour but are also motivated to act so that these interpretations become applicable to us, the reasons we have for acting in the ways we do become causally relevant for our behaviour. Our self-concepts are developed through our responses to our environment but they also play a role in shaping these responses.

The claim that our reasons for acting reflected in our judgments are causally active in our behaviour and that our self-concepts can be used in guiding our actions can lead to some thorny issues, however. Consider the aforementioned studies that hint towards the limits of our conscious awareness of the factors that lead to our behaviour. In light of these studies, it seems that we need to answer certain pressing questions if we don't want our account of human agency to be found wanting. Is the fact that our awareness of these factors is limited compatible with an account according to which we actively regulate our actions in accordance to our self-concepts? More generally, how is it that human beings, despite all the different factors influencing their perception of their environment and the way they respond to it, still maintain the capacity to respond in ways expressing their active self-control? In what sense can we be said to be in control of our actions, in a way that implies that we are the authoritative source of these actions and that we can be held responsible for them, when our actions are the product of a complex interplay of motivating factors that are not always subject to our conscious awareness? Before we attempt to provide an answer to these questions, we should examine in more detail some of the ways in which our conscious control over our actions has been shown to be limited.

The limits of conscious awareness and control

As a working definition, for the purposes of this discussion I'll take conscious processes to be effortful exercises of control that the agent can report as having occurred when asked[90]. An example of consciously controlled behaviour might be

[90] I say a working definition as I'm using it to draw an initial distinction between conscious and unconscious processes that will be useful for my following discussion. However, my aim is not to claim that effort and awareness, as I discuss them in the main text, are necessary and sufficient conditions for all forms of conscious control. Lack of effort, for example, might be less essential for a conscious process than lack of awareness (certain skills that come to require very little effort to perform might still be conscious in a sense, since their performers might still be aware of the exact

that of someone learning a swimming technique for the first time. In that case, the swimmer has to exert considerable effort in applying her knowledge of what the technique entails to her behaviour. She has to be aware of the various simple movements entailed by the specific swimming technique and she has to monitor and control her behaviour so that it manifests her knowledge of how to swim in a specific style. Another example of conscious processing might be that of an agent who actively reminds himself of his resolution to stop drinking after four glasses of wine when trying to decide whether to order another drink from the bar. Examples of this nature seem to require the agent's being aware of certain factors (the resolution to stop drinking after reaching a self-imposed limit or the knowledge of what a specific swimming technique entails) and to form an intention to behave in accordance to these factors. In implementing this intention, the agent has to monitor his on-going behaviour and to make sure that it doesn't divert from the behaviour he intends to engage in. Furthermore, these cases seem to entail that agents who consciously control their behaviour are aware of the main motivating factors that led to their actions and are able to report them accurately. When deciding not to order another drink, for example, the agent might report that it was his self-imposed limit on drinking that motivated him to refrain from getting another glass of wine.

Conscious processes, then, seem to be effortful and are subject to the agent's awareness, which makes them easy to report. What about unconscious processes? A way to draw a line between conscious and unconscious processing is to simply argue that any process that does not require effort or is subject to the agent's awareness so that it becomes easy to report is not conscious. A first way to approximate what it

movements performed during the exercise of these skills). In the main text, I discuss complex skills as unconscious, since they do not seem to require the kind of effort and awareness exhibited in the more straightforwardly conscious activities that I present (activities such as learning a skill for the first time). A different interpretation of various complex skills (such as dancing or driving) might instead treat them as conscious activities that require a different degree of conscious control than the conscious activities I present here.

For the purposes of the moral I intend to draw in this chapter, this distinction will not make a difference. As will become evident by this chapter's conclusion, even if only the activities requiring the direct, effortful conscious control described in the text are properly viewed as being conscious, and the rest of our behaviour is largely shaped by unconscious mechanisms, this would still not entail that we lack the capacity to act as authoritative self-controlled agents in all but the rarest cases involving direct conscious control. Furthermore, as I conclude, the kind of effortful direct conscious control I discuss might still have a role to play in our manifestations of our capacity to act as agents. I wish to thank Victoria McGeer for suggesting that I should clarify this point.

means to be unconscious is to argue that the operation of an unconscious process lies beyond the agent's awareness (so that the agent is unable to accurately report the influence it has on his behaviour) and is effortless, in the sense that it does not require the agent's active attention to, or monitoring of, his on-going behaviour in order to guide it in various ways. Unconscious processes are also frequently referred to as "automatic" in the relevant literature, since they seem to involve an agent's immediate response to his environment which is not mediated by any kind of conscious effort on his part.

John Bargh and Tanya Chartrand (1999a), for instance, distinguish between conscious and unconscious processes appealing to differences in effort, awareness and automaticity[91]. This is what they have to say on how we can distinguish conscious from unconscious processes:

> "The defining features of what we are referring to as a *conscious* process have remained consistent and stable for over 100 years […..]: These are mental acts of which we are aware, that we intend (i.e. that we start by an act of will), that require effort, and that we can control (i.e. we can stop them and go on to something else if we choose….). In contrast, there has been no consensus on the features of a single form of *automatic* process [….]." (Bargh and Chartrand, 1999a, p.463)

Despite the fact that it is hard to identify one single form of unconscious process, Bargh and Chartrand go on to argue that processes that are not conscious have been found to be "similar only in that they do not possess all of the defining features of a conscious process".(Bargh and Chartrand, 1999a, p. 463) A rudimentary distinction between these two kinds of mental processing is made possible, using the features commonly identified with conscious control and drawing conclusions regarding the type of process identified, based on whether these features are present or absent in a given case. Even with this distinction in mind, it is important to recognize that this is still only scratching the surface. As Bargh and Chartrand note, the nature of the influences on human cognition and behaviour that don't have the features commonly associated with conscious control is complicated enough to lead research into several diverging directions.

[91] See John A. Bargh and Tanya L. Chartrand, 1999a, "The Unbearable Automaticity of Being". *American Psychologist*, 54, pp. 462-479.

Well-honed skills seem to be one example of unconscious processing. Going back to the example of the swimmer, she might have to exercise significant conscious control in order to learn how to engage in a swimming technique, monitoring even very simple movements that have to be integrated in a complex behavioural display. Once she's mastered the technique, she doesn't need to exercise that kind of degree of conscious control on her behaviour and no longer needs to monitor every individual movement. This is plausibly true for a variety of complex actions that first require considerable conscious effort in order to be manifested in one's behaviour, from driving to dancing to playing a musical instrument. Having to consciously control every aspect of these complex activities would be disastrous in many circumstances and behaviour guided by unconscious processes is generally much faster and more efficient than behaviour that requires constant conscious monitoring (compare habitually tying one's shoelaces with trying to focus on every motion of the wrists and fingers while doing so).

Another type of unconscious process can be found in Bargh and his co-authors' (2001) research in non-conscious goal-processing[92]. As they present it, non-conscious goal-processing involves goals that "can be triggered outside of awareness and then run to completion, attaining desired outcomes."(Bargh et al. 2001, p. 1014). According to them,

> "[n]o conscious intervention, act of will, or guidance is needed for this form of goal pursuit. [N]onconsciously activated goals will cause the same attention to and processing of goal-relevant environmental information and show the same qualities of persistence over time toward the desired end state, and of overcoming obstacles in the way, as will consciously set goals." (Bargh et al. 2001, pp. 1014-1015)[93]

As indicated in Bargh et al's experiments, this kind of goal-processing can occur in cases in which subjects are neither consciously aware of the operation of certain

[92] See John A. Bargh, Peter M. Gollwitzer, Annette Lee-Chai, Kimberly Barndollar and Roman Trotschel (2001), "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals". *Journal of Personality and Social Psychology*, 81, 1014-1027.

[93] See also Bargh et al. 2001, p. 1014: "We postulate that mental representations of goals can become activated without an act of conscious will, such that subsequent behaviour is then guided by these goals within the situational context faced by the individual. In other words, just as most other areas of psychology recognize the nonconscious activation of mental representations, so too is it possible that goal representations do not need always to be put into motion by an act of conscious choice."

goals in their behaviour nor of their activation as a response to the features of their situation.

The main experimental set-ups supporting this hypothesis involved unconsciously priming (providing stimuli of which the subjects are not consciously aware) the participants with a certain goal and observing the effects of that goal in action. In Bargh et al's main experiment, the goal was for the participants to perform as well as they could in tackling a task provided for them, which consisted in finding as many words as possible in a word-search puzzle within ten minutes. In preparation for this task, the subjects were asked to complete a different word-search puzzle, with half of the participants having to find words related to success (e.g. achieve, win), and the other half having to find words that were neutral in meaning (e.g. river, hat). It was discovered that the subjects in the group primed with the words relating to success ended up doing better in the subsequent main word search task than the subjects who received no such priming. The interpretation provided by Bargh and his co-workers for these results was that the subjects who prepared for the main word search puzzle by finding words relating to success had the goal of performing well unconsciously activated in them, and that this goal operated unconsciously in these subjects' behaviour by leading them to perform well in the task.

Bargh and his co-authors also ran two similar experimental scenarios in order to examine the activation and operation of unconscious goals in greater detail. Both subsequent experiments also had a group of subjects primed with the goal of performing well and another group receiving no such priming. However, in one of these scenarios all subjects (who were recorded by a camera) also received, after two minutes had passed, instructions through an intercom to stop trying to find words in the main word-search puzzle. In the second scenario subjects were interrupted from the main task and subsequently given a choice to either resume the same task or switch to a different activity that did not involve the possibility of performing better or worse (rating cartoons based on how funny they were). In both scenarios, the group primed with the goal of performing well had the greatest percentage of subjects deciding to briefly persist in the main task even after instructed to stop and to resume working on this task after given the choice to switch to a different one. This led Bargh and his co-authors to argue that subjects that were primed with the goal of performing well not only acted based on that goal, but also acted in a very similar manner to that in which individuals having conscious goals guiding their behaviour act. That is because they resumed behaviour relevant to their primed goal if that behaviour had been interrupted and they were also able to overcome obstacles they faced while pursuing this goal.

However, the results obtained from the aforementioned set-ups might still not be sufficient for drawing a strong distinction between conscious and unconscious goals, since as admitted by the experimenters themselves their settings did not exclude the possibility that the only way unconsciously primed goals could play a role in their subjects' behaviour was if these subjects also had also received some kind of conscious instruction to perform in a specific way. The worry is that the subjects primed with the goal of performing well might have already had a conscious goal of doing well in the tasks set for them, activated by the explicit instructions they received, and that perhaps having such a conscious goal was a necessary prerequisite for having a similar unconscious goal operating in their behaviour. And one could argue that all the results of these experiments show is that the subjects that were primed with words relating to good performance did better than the subjects that were presented with words that were neutral in meaning because for the first group of subjects the unconscious goal of performing well was added as an extra motive to their already on-going conscious goal of finding as many words as possible in the main task. But does this show that unconscious goals can not only operate but also be activated unconsciously?

In response to this objection, Bargh and his co-workers conducted an experiment specifically designed to test whether subjects that were not consciously instructed to pursue a specific goal would still pursue it after being primed with words relating to it. In this instance, the relevant goal was that of cooperation. All subjects had to take part in a resource-management game against a simulated opponent in which they were both fishing from a pool with a limited number of fish. Part of the instructions all subjects received was that if the pool drained below a certain number of fish, both players would have to return all their resources to the pool. This game made different strategies available to all subjects, who were able to either cooperate with their opponent by returning a number of their fish to the pool in order for it not to drain too quickly, or compete against their opponent by attempting to gather as many fish as possible without refilling the pool, or use a combination of competition/cooperation However, half of these subjects were also explicitly instructed to cooperate with the other player as much as possible, and so to refrain from using strategies which did not involve any cooperation. The rest of the subjects were not given any explicit instruction on how to play the game.

Furthermore, some subjects from both groups were also primed with the unconscious goal to cooperate by having to solve a sentence-construction puzzle that consisted of words relating to cooperation, with the rest of the subjects having to complete a similar puzzle containing only neutral words. This puzzle was designed by the experimenters to prime the goal of cooperation on subjects from both the group that received an explicit instruction to cooperate and the group that was not instructed to play the game in any specific way. As such, there were four groups examined in the experiment, which consisted of subjects that were not consciously instructed but were unconsciously primed to cooperate (no conscious instruction/ unconscious priming group), subjects that were consciously instructed but not unconsciously primed to cooperate (conscious instruction/ no unconscious priming group), subjects that were both consciously instructed and unconsciously primed to cooperate (conscious instruction/unconscious priming group) and subjects that were neither consciously instructed nor consciously primed to cooperate (no conscious instruction/ no unconscious priming group).

The unsurprising result of this study was that the largest amount of cooperation was shown by the subjects in the conscious instruction/unconscious priming group and the least amount of cooperation was demonstrated by the subjects in the no conscious instruction/ no unconscious priming group. The more surprising and interesting result was that the subjects in the no conscious instruction/ unconscious priming group cooperated more than the subjects in the no conscious instruction/no conscious priming group and to a similar extent to both the subjects in the conscious instruction/ no unconscious priming group and the subjects in the conscious instruction/ unconscious priming group. The similar amount of cooperation shown by the subjects who were only primed with the unconscious goal to cooperate to the subjects in groups that were also explicitly instructed to cooperate is used by Bargh and his co-authors to support their previous hypothesis that unconscious goals can not only operate unconsciously, but also be activated unconsciously. Their results lend further plausibility to the claim that unconscious goals do not seem to always require corresponding conscious goals for their activation or subsequent operation, as the worry we previously examined would have it. According to this view, unconscious goals can be operative in one's behaviour without one necessarily having any conscious awareness of pursuing these goals and as previously argued, these goals can also guide one's behaviour in a similar way to conscious goals, persisting both in the face of encountered obstacles to their fulfilment and in the face of temporary disruption of their operation.

At this stage in our discussion, we can more convincingly claim that our behaviour is often guided by unconscious processes, some of which are activated unconsciously through making certain associations between perceived stimuli and corresponding courses of action. One such process consists in the activation of goals, such as the goal to cooperate with others, when encountering a context in which such a goal is applicable. Goals can be ones that have been regularly consciously pursued to the point where they are automatically activated when encountering a similar situation, but there is also the possibility that they are activated unconsciously without necessarily having been consciously pursued in the past (as Bargh et al's experiment involving cooperative behaviour indicates, for example). As further examination will reveal, there are other motivating factors that can influence our behaviour in similar ways, through being activated directly within various contexts we encounter wherein these factors become salient. More specifically, in addition to processes such as well-rehearsed skills and non-conscious goals being activated as a response to encountering certain features of the environment with which they have come to be associated, there is evidence in the empirical literature that other ways in which we respond unconsciously to our surroundings consist in the activation of stereotypes and the activation of the tendency to imitate perceived behaviour.

Taking stereotypes first, they are generally taken to be cognitive structures that consist of general features that are used in categorizing the objects of our social environment. Since we spend so much time interacting with one another, the most obvious targets for stereotyping are other people. We have developed a variety of stereotypes dividing people into groups consisting of features that are commonly associated with members of these groups. What makes the difference among groups could be anything from race, gender and religion to hair colour, profession or tastes in music. Different stereotypes are associated with Blacks, Whites, Asians, the elderly, rockers, blondes or fans of comic books, for example, with different general traits and characteristics associated with each group (forgetfulness with the elderly, long hair with rockers etc.). The typical problem with stereotypes is that they are often based on gross generalizations that are part of our cultural heritage and depend on antiquated and false notions of a given group's common traits. Stereotypes then, for everything they get right, might gloss over a lot of important ways in which people are different or similar and subsequently fail to do justice to the broad spectrum of human experience and variability. The stereotype referring to black people is a classic case of stereotypes gone bad, with traits such as laziness and hostility being associated with it. Using such a stereotype when interacting with others is bound to be problematic, since treating all members of a given group as lazy and hostile is more than likely not the best path to building a mutually respectful relationship with them.

Yet, as research into the nature and function of stereotypes has shown, these cognitive structures can be activated without our conscious awareness of their activation and have an effect on our behaviour that does not seem to depend on any kind of conscious monitoring and control on our part, much like in the case of the non-conscious goals examined above[94]. Simply perceiving some of the common characteristics associated with a given stereotype can activate that stereotype which, in turn, can have various effects on the way we behave, from affecting our evaluations of our surroundings to the actions we take. Furthermore, these effects might clash with our conscious attitudes on the subject. One theorist who has explored the contrast between conscious and unconscious factors when it comes to the function of stereotypes in our behaviour is Patricia G. Devine[95].

[94] See e.g. A.G. Greenwald and M.R. Banaji. 1995, "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes", *Psychological Review* 102(1), pp. 4-27.

[95] See Patricia G. Devine, 1989, "Stereotypes and Prejudice: Their Automatic and Controlled Components". *Journal of Personality and Social Psychology*, 56, pp. 5–18.

Devine's experiments were designed to test for any possible differences made by the level of conscious prejudice against black people displayed by her subjects, when it came to the activation and operation of stereotypes in their behaviour. In one experiment, Devine found that both high and low-prejudice subjects[96] had knowledge of the common features associated with the stereotype of Blacks. This result plausibly advocates against the possibility that low-prejudice subjects are not aware of the stereotype used by high-prejudice subjects. Devine's second experiment tested the effects unconscious priming of the stereotype of Blacks had on the behaviour of the subjects, when these subjects weren't in a position to consciously inhibit that effect. The experiment consisted in all subjects being unconsciously primed with various words, some of which relating to the stereotype referring to Blacks. The participants of this study consisted of both low and high-prejudice subjects. Half these subjects were primed with a list of words the majority of which were stereotype-related (80% of the list of 100 words) and the rest of the subjects were primed with a list of words the minority of which were stereotype-related (20% of the list of 100 words). All subjects were then asked to interpret a passage describing a person's ambiguous behaviour.

The hypothesis of her experiment was that, even though the stereotype-related words subjects were primed with were not directly referring to the trait of hostility, they were related to the overall stereotype used for black people (words such as ghetto, jazz and slavery). This would have as a consequence the activation of the respective stereotype (for the subjects belonging to the group exposed to the list of words the majority of which were stereotype-related) and the trait of hostility would still be activated for these subjects since it is part of the stereotype referring to black people. These subjects were then hypothesized to interpret the ambiguous behaviour as more hostile than the subjects that were primed with the least stereotype-related words. According to this hypothesis, the subjects would then interpret the ambiguous behaviour as more hostile without having any conscious awareness of being influenced by the priming of the respective stereotype. If that was the case, there

[96] Subjects were classified as high and low-prejudice after completing the seven-item Modern Racism Scale, which "is designed to measure subjects' anti-Black attitudes…[and] has proven useful in predicting a variety of behaviors including voting patterns and reactions to busing." (Devine, 1989, p. 7).

wouldn't be any significant difference between low and high-prejudiced subjects when it came to the unconscious activation and operation of stereotypes in their behaviour, when these subjects weren't in a position to consciously monitor and control this operation. Indeed, this hypothesis was corroborated in the results of the study, with subjects primed with the most stereotype-related words evaluating the ambiguous behaviour as more hostile than the rest of the subjects, regardless of their conscious level of prejudice. These results are supportive of the view that stereotypes can be unconsciously activated and have an effect on our behaviour even if they clash with our conscious attitudes towards the targets of these stereotypes. Even though in Devine's studies, this effect was limited to the subjects' evaluations of an ambiguous description of a behavioural display, the activation of stereotypes has been shown by further research to have a greater variety of effects than that.

One such effect consists in the enactment of "self-fulfilling prophecies" (Bargh and Chartrand, 1999a, p. 467). Perceiving features identified with a certain stereotype might lead one to use that stereotype in responding to the bearers of these features, because of that stereotype's unconscious activation and operation in one's behaviour. This can also provoke a similar response to the targets of the stereotype, by motivating them to adjust their behaviour so that it matches the expectations of the stereotype's user. In a study cited in Bargh and Chartrand (1999a), a visual task which had subjects observing a computer screen was used in order to subliminally present photographs of African American faces to some of these subjects[97]. This subliminal presentation's function was to unconsciously prime, in the subjects exposed to it, the Blacks stereotype, which as we've seen includes the trait of hostility. All subjects went on to participate in a two-player game, in which they took turns in trying to make their partners guess a specific word by offering various clues.

The result was that the subjects primed with the African American faces were more hostile to their partners than subjects not primed as such[98]. This led to the primed

[97] See Bargh and Chartrand, 1999a, p.467. The study described is M. Chen, and J.A. Bargh, 1997, "Nonconscious Behavioral Confirmation Processes: The Self-Fulfilling Consequences of Automatic Stereotype Activation", *Journal of Experimental Social Psychology*, 33, pp.541-560.

[98] For example, these subjects displayed hostile behaviour through the tone of voice they used and the level of annoyance and frustration they demonstrated when their partners offered wrong guesses. It is possible that the way the Blacks stereotype caused these subjects to display this kind of hostile behaviour towards their partners was by making them more likely to perceive their partners'

participants' partners adjusting their behaviour to match their partners' behaviour, so that they were also perceived as hostile by the primed participants. Chen and Bargh (1997) argued that, in the case of this study, the unconscious activation and operation of the trait of hostility through the Blacks stereotype played a self-fulfilling role, since the primed participants unconsciously manifested this trait in their behaviour and inadvertently caused their partners to respond in kind. The demonstration of such an effect is fascinating in its ramifications, since it's not too hard to imagine potential scenarios where interacting agents are locked in a vicious circle because of the effects of unconscious stereotyping in the ways they respond to one another.

This last example is also useful as a demonstration of our innate tendency to imitate each other's behaviour. A lot more is understood about this tendency now, with research in social cognition and neuroscience especially providing a basis for exploring the nature of our drive to imitate perceived behaviour. Susan Hurley, who has devoted some of her research to the subject, identifies different capacities that would not all constitute what she refers to as "full-fledged imitation" (Hurley, 2006, 7)[99]. For her, true imitation consists in copying the entirety of an observed action, from the means observed to achieve its result to the end it achieves. Other copying processes she identifies are "emulation" (achieving an observed result by trial and error learning), "response priming" (repeating certain movements without using them to achieve any particular goal) and "stimulus enhancement" (an action's drawing the observer's attention to something that triggers an innate response). As she makes clear, out of those copying processes true imitation seems to be the rarest, since it's

behaviour (e.g. the way their partners offered their guesses) as hostile, and so more likely to respond with hostility in turn. (I am indebted to Victoria McGeer for suggesting this interpretation to me).

Although Chen and Bargh (1997) do not seem to exclude this interpretation of their subjects' behaviour, they are less interested in the way the subjects subliminally presented with pictures of African American faces evaluate the behaviour of their partners, and more in the more overt behavioural effects that this subliminal presentation has on them (e.g. raised tone of voice, increased annoyance at wrong guesses). What is important, for the purposes of their study, is that however the initial hostile behaviour is initiated through the activation and operation of the Blacks stereotype in their subjects, it leads to the enactment of self-fulfilling prophecies. That is because the activation and operation of the Blacks stereotype leads to those subjects' partners responding in turn in a hostile manner, and thus to the confirmation of the initial expectations of hostility activated in the subjects that were primed with African American faces.

[99] See Susan Hurley, 2006, "Bypassing Conscious Control: Unconscious Imitation, Media Violence, and Freedom of Speech", in S. Pockett, W. P. Banks & S. Gallagher (eds.), *Does Consciousness Cause Behavior? ,* MIT Press.

mostly been found in humans and very few other animals, such as apes. What is more, the definition of true imitation as the copying of both goals and the means to achieve these goals is subject to considerable debate, as Hurley recognizes. Despite this complexity, I think Hurley's discussion can give us a first bearing on the functions of the tendency to copy perceived behaviour

For this tendency to have evolved, it seems that it must have certain advantages for our survival. What are these advantages? Hurley notes that for one, it allows for novel solutions to problems to be transmitted to others of one's kind not by biological but by cultural evolution. Imitators can assimilate adaptive behaviours which are not part of their genetic inheritance and thus preserve solutions to encountered obstacles that might not have been preserved if only a select few stumbled upon them by trial and error alone. For example, a creature with the capacity to imitate may observe its parents avoiding danger in an initially novel way and subsequently copy their behaviour, transmitting it to others along the way and maintaining this behaviour's importance for its species' survival.

Another advantage conferred by the tendency to imitate observed behaviour is that it allows for acting individuals to better interact with one another. In our case, what Chartrand and Bargh have called a "chameleon effect" (Chartrand and Bargh, 1999b) has been observed, which is postulated by these authors as facilitating our interactions with one another by allowing us to better fit our actions to the social contexts we encounter and to increase likeability among interacting individuals[100]. Subjects of their studies were found to unconsciously copy the mannerisms and posture of individuals they were interacting with, even when they had no conscious goal to get along with whomever was working in a task with them. The task chosen for these subjects was specifically tailored so that the subjects would not develop a conscious goal to facilitate their interaction with the confederate who was chosen as their partner, since it did not require any specific interaction between subjects and confederates other that individually working on the same task with minimal eye-contact. The fact that under such conditions the subjects copied the confederates'

[100] See John A. Bargh and Tanya L. Chartrand (1999b), "The Chameleon Effect: The Perception-Behavior Link and Social Interaction". *Journal of Personality and Social Psychology*, 76, pp. 893-910.

mannerisms supports the view that the subjects were manifesting an unconscious tendency to imitate their partner's behaviour, without necessarily doing so because of a conscious goal to facilitate their interaction with their partner. This kind of unconscious imitation was hypothesized to make interacting agents more positively predisposed towards one another and as such increase the chances agents have to fit in a given social context and to smoothly interact with others. This hypothesis is also supported by Chartrand and Bargh's (1999b) finding that confederates were rated as more likeable by subjects, after working in a common task, when they subtly imitated these subjects' posture and various mannerisms.

The dark side of automaticity

Up to this point, we have examined various unconscious effects our surroundings have on our actions by arguing for the existence of processes that consist in the unconscious activation and operation of cognitive structures such as goals and stereotypes and the tendency to imitate perceived behaviour. What these processes have in common is that they depend on a link between perception and action, so that our perception of certain features activates certain representations that are used in guiding our actions. A theory that is popular among theorists working on the nature of such processes and that can be used to explain the route from perception to action is the so-called "ideomotor" theory, according to which perception and action are inextricably linked, since they both share the same underlying mechanisms[101]. An exciting discovery in neuroscience pertaining to this theory involves the existence of "mirror neurons" which seem to be activated both when one is perceiving another's action and when one is acting the same way oneself[102]. The existence of these features has sparked significant debate which is not in the scope of this chapter to discuss, but it can serve as an example of the fact that support for a direct link between perception and action can be found in neuroscientific studies as well.

Thus unconscious mechanisms guiding human behaviour are already shown to be complicated enough to rival conscious control of one's behaviour. When I say rival

[101] See e.g. W. Prinz, 1990, "A Common Coding Approach to Perception and Action", in O. Neumann and W. Prinz (eds.) *Relations between Perception and Action,* Berlin: Springer, pp. 167-201 and W. Prinz, 2005, "An Ideomotor Approach to Imitation", in S. Hurley and N Chater (eds.) *Perspectives on Imitation: From Neuroscience to Social Science* (Vol. 1), Cambridge, MA: MIT Press, pp. 141-157.

[102] See e.g. G. Gallese, L. Fadiga, L. Fogassi and G. Rizzolatti, 1996, "Action Recognition in the Premotor Cortex", *Brain* 119(2), pp. 593-609 and G.Rizzolatti and L. Craighero, 2004, "The Mirror-Neuron System", *Annual Review of Neuroscience* 27, pp. 169-192.

conscious control, I mean that these processes can guide our behaviour in highly efficient ways that don't seem to depend on conscious monitoring of the on-going behavioural display. For instance, as we've seen that the studies conducted by Bargh and his co-workers indicate, the pursuit of an unconsciously activated goal can be resumed after being interrupted and individuals pursuing these goals can overcome obstacles to their attainment, despite not being consciously aware of having these goals. Does this have to be a downbeat conclusion? Not necessarily.

In the case of complex practiced activities, having the ability to train one's skills so that they operate unconsciously as a response to the appropriate context is generally viewed as enhancing the efficiency of one's actions. The smooth exercise of many complex skills seems possible only when one is not consciously monitoring every aspect of their operation. Furthermore, unconscious processes that depend on a link from perception to action (the activation and operation of non-conscious goals and stereotypes, imitative behaviour), which consist in the operation of cognitive structures activated through perceived features that are associated with them, can be construed as beneficial for our survival and social integration. One can argue that responding to a complex social environment becomes much easier through these unconscious processes, since they lessen the amount of effort we need to consciously exert on our actions. Bargh and Chartrand (1999a), for instance, frequently note this positive aspect of non-consciously responding to one's environment:

> "And so, the evaluations we've made in the past are now made for us and predispose us to behave in consistent ways; the goals we have pursued in the past now become active and guide our behaviour in pursuit of the goal in relevant situations; and our perceptions of the emotional and behavioral reactions of others make us tend to respond in the same way, establishing bonds of rapport and liking in a natural and effortless way. Thus, the "automaticity of being" is far from the negative and maladaptive caricature drawn by humanistically oriented writers [.....]; rather, these processes are in our service and best interests- and in an intimate, knowing way at that. They are, if anything, "mental butlers" who know our tendencies and preferences so well that they anticipate and take care of them for us, without having to be asked." (Bargh and Chartrand, 1999a, p. 476)

If we treat most of our unconscious responses to our environment in this manner, then we might be able to argue that the unconscious processes providing a direct link from our surroundings to our actions don't have as a consequence the diminishment of our capacity to express our agency in our actions. On the contrary, these processes can enhance our self-control, seeing as we can engage in activities expressing our agency in a far more efficient and effortless manner through the operation of non-conscious mechanisms. According to this interpretation of non-conscious processing, the fear that the pervasive influence of non-conscious factors on human behaviour means that human beings are driven by external forces beyond their control is unjustified since these processes can be understood as facilitating our expressions of agency in our actions, not as an impediment to our agentive status.

A potentially serious problem with this view is that it seems to rest on the explicit assumption that most of such unconscious processes were subject to one's conscious awareness and control at one time and that they became automated due to a repeated performance of a certain kind (e.g. stereotypes are automatically activated because of one's repeated conscious association in the past of certain features with certain groups of people). This brings in the notion of conscious control, since it seems that for the unconscious processes to facilitate one's behaviour and one's attainment of one's goals there ought to be a kind of conscious endorsement of this behaviour and these goals. It seems to me that there is some ambiguity when discussing unconscious mechanisms and goals. For example, in the processes that Bargh and Chartrand (1999a) describe as "mental butlers having our preferences at heart," what makes those preferences and interests, "our" interests? When one is pursuing a goal unconsciously, what makes this goal "his" goal?

The problem here is that as has already been noted, there is also the possibility for goals to be entirely unconscious, in the sense that they have been regularly and unconsciously activated and pursued in similar situations by an individual. That is, it seems that the aforementioned research is consistent with a view according to which goals can be activated unconsciously without there ever having been some conscious association between a certain context one is in and a certain goal related to that context. A person may routinely behave in an aggressive way to certain groups of people because of his unconsciously associating common features of these groups with certain negative stereotypes he possesses. Furthermore, this person might have acquired these stereotypes in a similar way to acquiring unconscious goals in Bargh et al's experiments, without ever being consciously aware of having them and of the effect they have on his behaviour.

In a similar way, chameleonic behaviour, whose potential advantages were briefly touched upon, might also constitute a threat to our self-control. Hurley's discussion of the effects violent entertainment has on its viewers provides an illustration of cases in which this kind of behaviour might threaten our ability to act as self-controlled individuals. According to her, combining research on our tendency to copy observed behavioural displays with research on the effects of violent media on our behaviour indicates that violent entertainment can affect our actions in ways of which we are not consciously aware. That is because the perception of aggressive behaviour can lead to the activation of cognitive structures used both for simulating and engaging in such behaviour which, if not inhibited, can lead the observer to also behave in an aggressive manner. One problem with such an effect is that this kind of inhibition might not be possible in cases involving observers with an underdeveloped or impaired capacity to inhibit their tendency to imitate the behaviour they observe. Children, and adults suffering from cognitive disorders, therefore, might not be able to adequately control the effects of viewing violent media. Even normal adults, as Hurley notes, might find it difficult to inhibit this effect if they are not consciously aware of its occurrence in the first place. So the case of being exposed to violent media could be a case in which our behaviour is affected in potentially harmful ways that we are unable to control because of our imitative tendency.

Another consideration that can reinforce the view that our capacity to act as agents is incompatible with the various effects our surroundings have on us is that we might frequently be mistaken as to what the actual motivating factors that led us to behave in a certain way were, since what we are consciously aware of might frequently be misleading us into making causal correlations that do not hold. Daniel Wegner and Thalia Wheatly, well aware of this problem, argue for an extreme proposal regarding the feeling of consciously willing one's actions[103]. According to them, conscious will is an illusion created by an acting individual's only being aware of the correlation between his conscious thoughts and his actions. What might be happening in this

[103] See Daniel.M. Wegner and Thalia Wheatley, 1999, "Apparent Mental Causation: Sources of the Experience of Will", *American Psychologist,* 54, pp.480-492.

case, as argued by Wegner and Wheatly, is that unconscious processes produce both the individual's actions and his thoughts about his actions. Because the subject that is acting is only aware of the path between his conscious thoughts about his actions and the actions themselves, he ends up feeling that his thoughts have caused his actions. But this feeling is mistaken in these authors' view, since what causes both thoughts and actions are mechanisms of which the acting individual is not aware[104].

When the various unconscious processes guiding one's behaviour are viewed this way, it becomes harder to see them as helpful "mental butlers" facilitating one's interaction with the environment by leading to the attainment of one's goals in a fast and efficient way. This is because the goals attained and the behaviours pursued in such cases are not necessarily goals or behaviours that would be consciously endorsed, were they subject to one's conscious awareness. Of course these goals and behaviours are still one's own, in a sense, since they are still products of the mechanisms that are part of one's entire physiological make-up. But in a more narrow sense these goals and behaviours are not one's own and they happen despite one's will, since were they to be consciously scrutinized they would not be endorsed by the individual pursuing them. This is where I think the reason a conjunction of Wegner and Wheatly's view on the illusion of conscious will and of the research on unconscious processes might seem disruptive for our agency becomes understandable. The worry that this research presents human beings as helpless puppets in a cosmic playhouse becomes more plausible in this case because it seems that there are unconscious processes guiding one's behaviour that are fast and efficient but which nevertheless do not promote one's own goals, in the narrow sense. The mental helpful butlers in this case seem to turn into mental invisible

[104] See also Benjamin Libet, 1985, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action", *Behavioral and Brain Sciences* 8, pp. 529-566, for a series of experiments that can be taken to support Wegner and Wheatly's account. Libet's goal in these experiments was to examine the correlation between his subjects' conscious awareness of intending to act in a certain way, their voluntary actions and the brain processes related to these actions. What Libet found was that there was significant cerebral activity related to the preparation of the motor movements that were part of these subjects' actions (which consisted in their flexing of their wrists or fingers) and that this activity preceded his subjects' conscious awareness of intending to act in a certain way (which was based on their reports of when they first became aware of their intentions to act). His interpretation of these findings was that we do not directly consciously initiate our voluntary actions, as these actions are initiated by brain processes we are not aware of, but we might have the ability to interrupt them before completion.

intruders inexorably enforcing their wishes on the goals and behaviours being pursued by the acting individuals.

As is already evident from the introduction to this chapter, I think that finding this grim prospect inevitable belies a confused understanding of our nature and function as complex creatures. Being in a position to examine our empirical nature in more detail can enable us to better appreciate the fact that we are enormously complex living organisms, both in terms of our constitution and in terms of our responses to our surroundings. An appreciation of this fact, however, should not lead us to endorse the aforementioned view. We do not lose our capacity to act as agents because of this complexity and we are not constantly driven by forces beyond our control. To the contrary, this complexity enables us to better act as agents by supporting our capacity to express ourselves in our actions in an authoritative manner. The fact that the subjects in the aforementioned studies have been shown to be particularly vulnerable to experimental manipulation and to have a limited conscious awareness of the springs of their behaviour, as well as limited conscious control over their actions, is not a contradiction to this standpoint. It can serve as a sombre reminder that we are flawed in many ways and that our ability to express our first-person perspective in our actions and to be in control of what we think and what we do is far from perfect. But an alarming “puppets in a cosmic playhouse” reaction is an unjustified leap from recognizing how malleable and complex our nature is to underestimating our ability to engage in self-controlled actions that express our individual viewpoints. The simple answer for why we can and do, in fact, act as agents, even in the face of explanations that strip our nature to its nuts and bolts, is this: we are social creatures that have been trained through their environment to express themselves in their actions and to expect each other to act as such. Elaborating this claim and its implications is my aim in the remainder of this chapter. A more nuanced understanding of the difference our peculiar social circumstances make for the kind of self-control we are able to exhibit in our actions can lead to the adoption of a more balanced perspective on the significance of the interplay between the conscious and unconscious motivating factors that are involved in our responses to our environment.

Social constraints on agency

The explanations given of the behaviour displayed by the subjects in the aforementioned empirical studies tend to lean towards mechanistic descriptions of these various individual behavioural displays. These explanations tend to focus on the particular combination of environmental effects and physical mechanisms that have led to the behavioural patterns in question. In the case of unconscious goal activation and pursuit, for example, giving the etiology of the behaviours displayed by the subjects in the relevant studies involved invoking the combination of the activation of unconscious processes, primed in a specific experimental setting, and the operation of these processes in producing the observed behavioural patterns. Subjects cooperated more and persisted more in overcoming encountered obstacles because of the operation of unconscious goals being activated through experimental manipulation. Something similar can be said for the rest of the examined studies. Subjects respond in a particular way to their circumstances because of the effect coming in contact with specific conditions has on them. These conditions have this effect on them because of their cognitive constitution's sensitivity to them. Perceiving someone's facial features can activate a corresponding stereotype, for example, or perceiving someone's gestures can activate the tendency to imitate them. These mechanistic descriptions serve as adequate explanations of the subjects' behaviour because they reveal its most plausible physical causes. And they can indeed be highly useful in enriching our understanding of the nature and functions of the various mechanisms subserving our actions.

The problems start when using these explanations to derive conclusions about whether the observed actions are the product of exercises of self-control and of the subjects' capacity to express their viewpoints in their actions. The assumption driving the use of these explanations in providing these answers is that we can identify whether agency has been expressed in one's actions by providing a description of their causes, in terms of the way one's internal structure is affected by one's circumstances. If we can't identify anything corresponding to our definition of agency in invoking the causes of one's actions, then it seems that there is no way to make sense of talking about one's expressing one's agency through the actions observed. Furthermore, if we use a more strict definition of what counts as an action, according to which something is an action if it is the product of an exercise of active self-control, then we might even deny that it makes sense to speak of "one's actions", since they don't fit our definition of an active expression of one's agency.

How can we identify an expression of agency in a description of the causes of one's behaviour? Well, that depends on our account of agency. Leaving aside the account I've been developing up to this chapter, which depends on a synthesis of mainly philosophical perspectives on mind and action, we can try to discern the simpler, not as theoretically mediated intuitions that seem to drive the assumption that discovering the actual causes of our behaviour significantly threatens our agency. The main intuition seems to be that there has to be a clearly identifiable mental act that we can identify as the expression of one's agency. This act, furthermore, needs to be conscious, in the sense that it needs to be something that one is aware of when one acts as an agent. If we use our initial distinction between conscious and unconscious processes, it seems that what we're looking for needs to be an act which one is aware of and which expresses one's control over one's actions, and not a direct response to one's environment that operates beyond one's conscious awareness. It seems that one needs to be consciously aware of the mechanisms leading to the actions one engages in and that, furthermore, one needs to consciously initiate and guide the operation of the various mechanisms leading to these actions.

This viewpoint is what Philip Pettit calls the "act-of-will picture," whose validity for serving as the basis of an account of what it means to act as an agent he intends to undermine[105]. In his view, explanations focusing on the physical causes of individual behavioural displays in order to draw conclusions on whether these displays are manifestations of active self-control are missing out on the difference our social environment makes. Pettit insists that the act-of will picture needs to be jettisoned in favour of a picture involving what he calls "agent-control," wherein agency is treated as a complex, emergent, social phenomenon. The main difference between the two pictures concerns the factors that are deemed relevant to ascertaining whether an action counts as a product of agency or not. Whereas in the act-of-will picture, what is deemed relevant is the individual's internal machinery, in the agent-control picture the individual's place in a society in which there is a common understanding and

[105] See Philip Pettit, 2007, "Neuroscience and Agent-Control", in D. Ross, D. Spurrett, H. Kincaid and L. G. Stephens, (eds.), *Distributed Cognition and the Will*, MIT, pp. 774-789.

expectation of what counts as a product of responsible agency is also taken into account. For a behavioural display to count as agent-controlled, the agent has to be in a position to justify it as his own action and to defend it against alternative explanations. The agent is expected to be in a position to provide reasons for acting in the ways he does and these reasons have to be sound enough so that they both enable him to see these actions as expressing his agency and are accepted by other agents that are trying to explain these actions. By offering reasons for his actions, the agent is put in a position of authority regarding his control over his own behaviour. This authority can be undermined by the agent's inability to integrate his behaviour with his reasons for acting with the result that his claims of self-knowledge and self-control might be disregarded in favour of alternative explanations.

How can the agent's authority be undermined? That depends, as Pettit argues, on the common standards of what counts as an agent-controlled action that are recognized by the agent and shared by the members of the agent's current society. The agent has to live up to these norms and his ability to own his actions and express his active viewpoint in them depends on his ability to respond to these norms, in a way that is satisfying both for himself and others. Whether an agent consciously initiates and controls the mechanisms leading to his actions by an act of will is not relevant, when adopting this view. What we should be looking at is the capacity agents have to provide reasons for their actions and to live up to these reasons, a capacity for which Pettit uses the names "conversability" and "orthonomy"(Pettit, 2007, p. 83). Conversability is meant to highlight the agent's ability to explain and defend his reasons for acting, while orthonomy is meant to emphasize the agent's ability to be guided by these reasons. Acting as an agent requires being conversable and orthonomous in these ways, so that it requires a capacity to provide, defend and be guided by one's reasons for acting. This capacity is presupposed by agents that expect one another to provide satisfactory explanations for why it is they act in the ways they do and its successful exercise is conducive to maintaining one's authority and control over one's actions.

How does this view fare in the face of the fact that human behaviour is the product of a multitude of complex processes that are frequently beyond conscious awareness? Pettit intends the agent-control picture to be compatible with advances in the sciences exploring mind and action. The fact that our cognitive complexity undermines a simple picture according to which agent-controlled behaviour is the product of a traceable conscious act should not be taken as an argument against attributions of agency, since we can think of agency on a different scale that is not dependent on tracing a simple conscious act as the original cause of behaviour that's under the agent's control. Our complex nature should be taken to support our capacity to act as agents not because of its subservience to conscious acts of will but because it supports our capacity to be guided by reasons for acting. As Pettit phrases this:

> "On the act-of-will picture, it is in virtue of the fact that an unfolding action is subject to my perceived or phenomenal control that it counts as agent-controlled. But, if the argument here is correct, it is not in virtue of being subject to that phenomenal control that the action is agent-controlled. Rather, it is in virtue of being agent-controlled-in virtue of being performed in the presence of a neurally supported capacity for conversability or orthonomy- that it has such a perceptual or phenomenal profile."(Pettit, 2007, p. 89)

Our levels of neural and cognitive complexity, therefore, have enabled us to achieve the kind of sophistication required to develop an understanding of what it means to act as a self-controlled individual and to be guided by this understanding. Another implication of this view is that we did not start out, as a species, biologically equipped with the ability to express ourselves in our actions in an authoritative manner. This ability has developed through our training to understand what it means to have reasons for acting in a certain way and to be able to proficiently engage in an exchange of such reasons in our interpersonal relationships. The collaborative social framework in which we develop our cognitive capacities uniquely shapes our responses to one another and to our circumstances and is required for the development of a capacity to act as agents that are able to justify and take responsibility for the actions they engage in[106].

The reason I find Pettit's view attractive should be obvious, given the account I've been developing in the preceding chapters. I am in general agreement with Pettit's argument that we should develop an account of agency as a complex, social

[106] See Pettit, 2007, p. 87: "A capacity like the capacity to be conversable or orthonomous is inevitably the product, not just of native makeup, but also of cultural development. We are not born responsible, any more than we are born free".

phenomenon that depends on an ability to make authoritative claims about our actions and justify these claims to ourselves and others through our behaviour. The capacity to recognize and live up to reasons for acting, reasons that can be shared by interacting agents to allow them to make sense of one another's behaviour, is crucial for the present account. But we need to be clear on what this capacity involves and on why it is compatible with the research on the various unconscious factors influencing our behaviour.

A danger with a view such as Pettit's is that it might seem to wilfully ignore the empirical evidence, instead of accommodating them. The claim that agency is a complex, social phenomenon and that focusing on the immediate causes of individual behaviour doesn't do justice to this phenomenon, so that we should shift our thinking on the matter from a restricted, narrow view of self-control to a broad understanding of it that encompasses normative considerations involving our constant collaboration in maintaining our society, is deeply important. It needs to be elaborated carefully, though, so that it avoids the accusation of sweeping the empirical facts under the carpet. This account of agency should be supplemented with an examination of how our common understanding of what it means to act as an agent influences our actions on the individual level. Furthermore, the nature of the reasons that are frequently evoked in Pettit's account should be elaborated on, so that their importance for an account of agency and the way they enable us to act as agents become clearer.

If our account of agency is not elaborated in these ways, it faces the danger that Pettit's view is facing, of being treated as an unsubstantiated philosopher's fiction borne out of a desire to defend our common practice of treating one another as responsible individuals that are in control of their minds and actions, even in the face of worrying empirical evidence. Pettit has already provided us with the general guidelines of avoiding this danger, by showing that we can be unaware of the various factors influencing our responses to our environment while still being able to recognize that our common practice of treating one another as agents is a fundamental aspect of our nature and not just an explanation that is only adequate until made redundant by advances in the understanding of our empirical nature. Sketching a clearer picture of the manner in which our social nature enables us to act as agents will involve understanding how our empirical nature changes as we develop the capacity to act as agents and applying this understanding to our discussion on the interplay between conscious and unconscious processes in the production of action.

Learning to play the role of the agent

The account expounded in the previous chapters and summarized in the introduction to this discussion can provide a useful framework for extending Pettit's argument on agent-control. Recall that the main idea in that account was that agency consists in the exercise of a special kind of self-regulation, which involves regulating one's actions in accordance to one's self-concept. The self-concept playing a guiding role in action is based on the judgments made by the agent that function as normative commitments for him because they motivate him to express the intentional states corresponding to these judgments in his actions. Furthermore, these judgments are filtered through the agent's folk-psychological understanding of himself and his environment, so that the intentional states expressed by them are subject to certain norms implicit in this folk-psychological understanding (e.g. the norm to avoid expressing contradictory intentional states) and can be expressed in a public language. The normative status of these judgments comes from the fact that they can be examined by both the agent making them and other agents he interacts with, so that they can be used in both justifying this agent's actions and creating expectations on what actions he will take, if his judgments are to be taken as genuine expressions of his understanding of himself and his environment. Therefore, as we've also touched upon, since these judgments can be used in both justifying and predicting the agent's actions, they can be taken as expressing the reasons that the agent has for acting in the ways he does.

In viewing agency as a social phenomenon, we started with Pettit's idea of the development of a capacity to be guided by reasons and to be adept in explaining and predicting one another's actions by reference to reasons for acting. Now we're in a position to elaborate on how this capacity evolves and what its exercise might involve. The judgments expressed in the agents' self-concepts play the role of the reasons for which these agents act, when these agents regulate their behaviour in accordance to the way the view themselves as responding to their circumstances. In the cases in which the intentional states manifested by the agents in their behaviour correspond to their judgments on how to respond to their environment, their actions can be made sense of by reference to these intentional states that are expressed in the judgments that are part of their self-concepts. Pettit's capacity towards orthonomy or conversability can be adapted in our current framework as a capacity to coherently manifest the intentional states expressed in one's judgments so that they fit one's self-concept.

Coherently expressing one's intentional states in one's actions involves regulating one's behaviour so that it manifests the dispositions expressed in one's normative judgments. I'm stressing this point so as to make clear that the agent's reasons can motivate the agent to act in ways consistent with those reasons. A crucial aspect of the account I'm using is the idea that one's self-concept is not simply a construct used to justify and predict one's actions, but is also used by the agent in orchestrating his various behavioural displays (which can range from thoughts on a subject to overt actions) in the shape of actions that make sense, to himself and to others, as an enactment of the intentional states expressed in the agent's self-concept. Looking at the issue in this way, Pettit's point that the capacity to act as an agent involves being guided by one's reasons, or what he calls orthonomy, can be rendered more intelligible. The agent's reasons guide his actions when the agent regulates his behaviour so that it fits his self-concept expressing the judgments playing the role of his reasons for acting.

The capacity to competently exchange reasons with other agents in order to maintain one's authoritative control over one's actions, or what Pettit calls conversability, can also fit in our current account. The way the agent is guided by his reasons for acting is dependent on the folk-psychological framework that determines whether the intentional states that are expressed in his judgments are intelligible, and on the agent's capacity to appropriately manifest his intentional states so that his judgments become accepted as a legitimate justification of his actions by agents that have a folk-psychological understanding of the nature of intentionality. The agent is expected to live up to his normative judgments so that he is understood as actively expressing his self-knowing contribution in his responses to his environment. As the intentional states expressed in these judgments are subject to the norms inherent in a folk-psychological understanding of intentionality, the agent is expected to respect these norms in living up to his self-concept. Consequently, the agent has to be able to manifest his intentional states in his actions without violating these norms to such an extent so that his authority is questioned. To go back to a previous example, if an agent judges that an apple is made of wax but proceeds to treat it as a real apple, his ability to be guided by his reasons for acting is questioned. If, furthermore, the agent tells everyone that wax is inedible but then takes a bite out of the wax apple, he might be viewed as expressing two incompatible intentional states in his actions or as not being able to accept the consequences of his attitudes, in which case his understanding of what it means to express these attitudes coherently is brought into question.

At its present stage, our account can already serve as a useful framework for comprehending the peculiarities of our social nature and the difference it makes for how we learn to express our agency in our actions. Before we further examine how these considerations bear on our discussion of the difference between conscious and unconscious processes and of the pervasive influence of unconscious influences on our actions, I think this account will benefit from drawing some parallels with the work of other authors investigating the difference our social nature makes for the nature and function of our cognitive capacities. There are several viewpoints that I think are interesting in this context: Daniel Dennett's focus on the difference a demand for reasons in a common language makes for our cognitive capacities[107]; J. David Velleman's comparison of the interactions among human agents that share a common understanding of intentionality with the interactions engaged in by a troupe of collaborating self-improvising actors[108]; John Bargh's (2006) answer to the question of how multiple parallel effects are triggered through seemingly simple priming procedures[109].

Dennett's work has already played an important role in shaping our current account. My preference for a view of distributed control and agency owes a lot to his arguments against views in which some kind of central controller of action is

[107] See especially Dennett, 2003.

[108] See especially Velleman, 2009.

[109] See John Bargh, 2006, "What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior", *European Journal of Social Psychology,* 36, pp. 147-168.

postulated in order to explain the unity of our actions. Also, explaining self-regulation as fitting one's self-concept to one's actions has been partly motivated by his insights on the capacity of highly complex self-organizing creatures to develop a self-concept that plays a crucial role in their actions. Dennett's work has also ranged from tackling questions on the nature of consciousness, to debates on free will and responsibility and on the development of moral agency. The latter aspect of his work I find especially relevant, in the present context, since Dennett's answers to the question of whether we can maintain our status as rational, self-controlled agents in the face of scientific progress revolve around noting the transformative effects our social nature has on our ability to express ourselves in our actions.

In "Freedom Evolves" Dennett expresses the following worry: "Aren't we learning from psychologists that we are *actually* a far cry from the rational agents we pretend to be?" (Dennett, 2003, p. 268). His response is that, actually, by pretending to be agents we make ourselves into agents. In a move similar to Pettit's, Dennett argues that we should consider our common practice of intentional interpretation of one another's actions in order to understand the nature of our capacity to act as responsible agents. In his view, having the ability to inquire into one another's thoughts and actions led to an increase in the sophistication of our capacity to monitor and control our behaviour. The development of a language in which we could express our intentional states and our reasons for acting, making them publically available objects of inquiry, was the decisive step in this process.

As Dennett's story goes, when that level of interaction was achieved,

> "people could *do things with words* that they could never do before, and the beauty of the whole development was that it *tended* to make those features of their complicated neighbours that they were most interested in adjusting readily accessible to adjustment from outside-even by somebody who knew nothing about the internal control system, the brain. These ancestors of ours discovered whole generative classes of behaviors for adjusting the behavior of others, and for monitoring and modulating, (and if need be, resisting) the reciprocal adjustment of their own behavioral control by these others." (Dennett, 2003, p. 249)

Dennett also makes clear that this increase in self-control depends on our capacity to offer reasons for our behaviour and expect others to do the same:

> "We human beings can not only do things when requested to do them; we can answer inquiries about what we are doing and why. We can engage in the practice of asking, and giving, reasons.
>
> It is this kind of asking, which we can also direct to ourselves, that creates the special category of voluntary actions that sets us apart". (Dennett. 2003, p. 251).

This view is compatible with our general theme in this chapter that being able to treat one's reasons for acting as an object of public inquiry motivates one to live up to these reasons by acting in ways that can be explained by reference to these reasons. Pretending to be an agent, in accordance with Dennett's story, involves trying to live up to the reasons that would make it possible for one to be treated as an agent. Trying to live up to these reasons, which can be expressed publicly and thus be used in evaluating one another's behaviour, leads to developing a degree of self-monitoring and self-control that is instrumental to being able to fit one's behaviour to these reasons. Through this ability to control one's behaviour so that it manifests one's reasons for acting, one gains authorship of one's actions because one is in a position to justify them by invoking the reasons for which they were performed.

A further relevant argument that appears in various guises in Dennett's writing, from "Consciousness Explained", to "Kinds of Minds" to "Freedom Evolves", is that we develop self-concepts which enable us to explain our actions as the expressions of a stable, coherent point of view. Dennett's understanding of the role of this construct in action has not always been clearly expressed throughout his writings (see for example, the relevant discussion in Chapter 3). This idea, however, is essential to providing a more complete reconstruction of his story of the development of rational agency. A self-concept is essential for being able to live up to one's reasons for acting, because, to put it in a ways closer to Dennett's terminology, the agent can tell stories about what he's doing and why, stories that fit the attitudes expressed in this concept. One of Dennett's more recent formulations of this familiar argument goes like this:

> "The acts and events you can tell us about, and the reasons for them, are yours because you made them-and because they made you. What you are is that agent whose life you can tell about. You can tell us, and you can tell yourself." (Dennett, 2003, p. 255)[110]

[110] For a more extensive use of the notion of fitting such a narrative to one's actions, see the previous chapter's discussion of Dennett and Velleman's views on the self as narrator.

The importance of a folk-psychological understanding of the way people think and act for our capacity to act as agents is also a theme in Dennett's work that has obvious parallels with our own account of agency. Dennett's "intentional stance" is the framework in which agents offer and exchange reasons for one another's actions. The intentional stance involves interpreting observed behaviour by referring to the existence of certain intentional states explaining these actions. Our use of this stance depends on our implicit folk-psychological understanding of human mind and action, so that any kind of attributions we make are filtered through this understanding. Seeing as having a reason for one's actions is understood as being able to manifest one's intentional states so that they express this reason, developing the capacity to live up to one's reasons, in Dennett's own work, seems to also depend on a folk-psychological understanding of what it means to express these reasons in one's actions.

Dennett often invokes a tale in which our common training in folk psychology led to the common demand for reasons, according to which we were initially responding to one another in a way that did not depend on any kind of understanding of what we were doing, but were subsequently (due to a combination of biological and cultural evolution) able to examine these responses themselves and the norms governing their use[111]. Dennett also frequently notes the significance of early development and of a child's interactions with its peers and its caretakers for developing the kind of understanding that underlies our common demand for reasons. This account, according to which the intentional stance grounds our common practices that enable us to express ourselves in our actions in novel ways, is one way to understand one of the prevalent themes in our own account (see especially Chapter 1); our training in folk-psychology enables us to understand what is expected of us when acting as agents and to subsequently act as such and expect others to do the same[112].

[111] See the previous chapter for McGeer and Pettit's use of a similar story and of Dennett's intentional stance in order to distinguish self-regulating from routinized intentional systems.

[112] A note on Dennett and consciousness: I agree with Andy Clark in his 2002 that, while Dennett seems to treat his story as a pre-requisite for understanding what it means for a creature to be conscious, this is a problematic conclusion that doesn't naturally follow from what Dennett argues. See Andy Clark, 2002, "That Special Something: Dennett on the Making of Minds and Selves", in A. Brook and D. Ross (eds.) *Daniel Dennett,* Cambridge University Press, p. 197:

Velleman, another author whose ideas have been essential to the development of this thesis, provides a useful framework for considering the importance of our exposure to social practices for our ability to authoritatively express ourselves in our actions[113]. In previous chapters, his work has enabled us to see how we can provide a reductive account of agency without inventing a mysterious agent that has to identify with certain motives or actions. Reductive accounts of agency usually fall prey to the problem of infinite regress, but as argued in Velleman's work, a motive for making sense of one's actions, or acting in accordance to reasons, is a plausible candidate for playing the functional role of the agent, as it is something the agent cannot dissociate himself from without losing his status as such. In this view, when we act as agents we are motivated to do so because of the need to make sense of our actions as expressing ourselves. We can understand our actions as such by viewing them as expressions of our reasons for acting.

Velleman's account of narrative control, furthermore, provides a plausible direction for the exploration of the idea that the narratives centred in our self-concepts (or, in Dennett's terms, the stories we tell about ourselves) can guide our actions. Velleman's relevant insight is that the motive towards making our actions intelligible by reference to the reasons behind them, which is constitutive of our agency, can also be viewed as a narrative module that creates an integrative whole from our actions, our self-concepts and our circumstances. This narrative module plays the function of fitting the stories we tell about ourselves to the actions that we perform and vice versa. In adapting this idea to our account, I have argued that we should avoid falling into the pitfall of making this narrative module a mini-agent, but instead argue that the agent himself acts as a narrator because of his motive to make his actions intelligible to himself and to others.

A metaphor that I have found particularly evocative and that is central to Velleman's thinking is that of the self-improvising actor, who has no set script but continuously improvises and enacts his role according to his circumstances. In his

"[For human agency, it surely *is* the practice of public, language-dependent, criticism and reflection that instills in us the kind of meta-reflective skills that Dennett…highlight[s]. The "cognitive bonus" that language confers thus seems central not just to the incremental learning of abstract concepts…but also to the emergence of morally responsive agency…There is still nothing here, however, which speaks to the rather bulky remainder of our matrix of mindfulness: the presence of qualitative consciousness and the potential for significant suffering."

[113] See especially Velleman, 2009.

(2009), Velleman extends this metaphor and his aforementioned insights to our interpersonal interactions. He argues that we don't only enact a role for ourselves but that we also need to present ourselves in a coherent manner to other self-improvisers. We not only need to make sense of our actions as expressing a coherent point of view but we also have to present these actions as such to others, in order to engage in coherent interactions. Because of this aspect of our nature, the enactment of our self-concept, or reasons for acting, depends on a mutual understanding of what it means to engage in intelligible enactments of this sort. Like a troupe of collaborating self-enactors, we depend on one another's cues for understanding the context in which we find ourselves and the actions that would be available within that context.

As I previously argued (see especially Chapter 2), Velleman's story seems to fit in many respects with an account in which our interactions based on a common understanding of intentionality lead to the development of our capacity to be guided by reasons. An interesting question concerning the motive for self-understanding is what the relationship between this aspect of our nature and our social interactions is. As I argue in Chapter 2, such a motive would seem to enable us to act as agents only within a folk-psychological social framework in which we can understand our actions as expressing our reasons for acting because they manifest our self-attributed intentional states. Velleman seems to think that the specific way in which our folk-psychological understanding develops and our social interactions take place depends on our innate drive to make sense of our actions as our own. My view is that we might indeed have an innate drive to make sense of ourselves as distinct from our surroundings but this drive only becomes a motive to understand our actions as the products of our agency at the stage where we have developed a common understanding of reason-guided action. If we interpret the interplay between an agentive drive towards self-understanding and the social interactions built around it in this way, Velleman's story seems to provide a useful metaphor for how we think and act.

In his (2006), John Bargh takes a step back from findings on the variety of effects observed in priming studies in order to assess the overall validity of these studies. We've discussed such effects in our own discussion, from the activation of stereotypes to the activation and operation of unconscious goals to cooperate or to perform well in solving puzzles. These effects are interesting in their own right but, as Bargh argues, there are certain problems pertaining to their general relevance to the way we think and act that need to be addressed. Otherwise, effects such as the influence of unconscious goal activation and pursuit on action face the danger of being treated as "psychological parlor tricks" (Bargh, 2006, p.150), obtained only in artificial situations that have very little to do with the way we normally behave.

A problem priming studies face is what Bargh refers to as the "generation problem" (ibid, 152). The central question is this: how is it that simple priming procedures, (subliminally flashing a word to a subject, for example), have such varied and simultaneous effects on action? One example from our discussion is priming subjects with the trait of hostility through the Blacks stereotype. We've seen that such priming can cause subjects to evaluate behaviour as more hostile (Devine 1989) but also behave in a more hostile manner themselves (Chen and Bargh 1997). As a further illustration, here's what Bargh has to say about the priming of a single stimulus:

> "[A] priming stimulus such as *generous* can be expected to (1) activate affectively similar but otherwise semantically unrelated material in memory […..]; (2) impressions and trait judgments of a target person who behaves in an ambiguously generous manner […]; (3) increase the likelihood of a generous behavior under general circumstances (e.g. being asked to donate to a charitable organization); (4) trigger altruistic motivations and goal pursuits […..]. Priming effects, it seems, come in packages- constellations or thematically related sets of effects." (ibid, pp. 152-153)

Which of these effects is discovered in a given study largely depends, as Bargh notes, on the different focus and interests on the part of the experimenters. Though in his own work, Bargh has devoted a big part of his research on exploring the activation and pursuit of goals of which the subjects are not aware, he not only recognizes that the same priming methods can simultaneously affect the subjects in different ways that are not limited to the activation and pursuit of such goals, but also that there might be different goals activated at the same time through encountering the same stimulus. In Bargh (2006), two studies are cited in which the same priming manipulations activate different sets of goals. In both studies, priming representations of subjects' familiars activated certain goals for the participants, but in the first study the experimenters found that goals affecting their subjects' behaviour were ones that these familiars had for the subjects, while in the other the goals which the subjects pursued when coming into contact with these familiars were found to be active in their behaviour[114].

What, exactly, is being primed in all these cases, if there are multiple aspects of the subjects' cognitive constitution and behaviour that are being affected by encountering the relevant stimuli? Bargh's answer to this is that what is being activated as a response to these stimuli is some kind of cluster of interconnected associations with them. Through this conclusion, Bargh is articulating what seems to be the subject of increasing common consensus among diverging disciplines:

> "Thus, several disparate areas of research and social thought lead us to the same conclusion: that one reason for the multiple parallel effects of our priming manipulations is that we might not be priming single concepts, but rather conceptual structures, whether they be called metaphors, roles, perspectives, or mindsets". (ibid, p. 158).

In his view these perspectives, or roles, are affecting different aspects of the way subjects perceive and respond to their environment, from their evaluations of others' behavioural displays, to the goals they pursue. Which perspective is activated at a given time depends on the circumstances subjects find themselves in and on the conceptual associations that are most relevant to these circumstances. In one case, for example, subjects might be primed with stimuli activating the perspective of a generous person, as in Bargh's previous example, which leads to a variety of different cognitive, emotional and behavioural effects.

The aspect of Bargh's view that makes it fit the context of this chapter is that he considers interpersonal interactions, especially those occurring during early stages in human development, as crucial for the emergence of these conceptual structures. The claim here is that from an early age, we come into contact with the way others experience and respond to the world and with the conceptual structures underlying these experiences and responses. Coming into contact with these perspectives shapes the way our own cognitive development, which also consists of forming conceptual associations that can be activated under their corresponding circumstances, proceeds.

[114] See Bargh, 2006, p.152. The studies cited are J.Y Shah and A.W. Kruglanski, 2003, "Automatic For The People: How Representations of Significant Others Implicitly Affect Goal Pursuit", *Journal of Personality and Social Psychology,* 84, pp.661-681 and G.M. Fitzsimons and J.A. Bargh, 2003, "Thinking of You: Nonconscious Pursuit of Interpersonal Goals Associated With Relationship Partners", *Journal of Personality and Social Psychology,* 84, pp. 148-164.

A way to put this that would also be consistent with Bargh's view is that we learn to adopt different perspectives, or roles, appropriate to our circumstances, through encountering the roles or perspectives that others manifest when encountering similar circumstances. Bargh demonstrates that support for this claim can be found in the work of authors coming from fields of study as diverse as philosophy of mind and developmental and political psychology[115].

[115] As examples, philosopher of mind Charles Fernyhough (see Fernyhough, 1996, "The Dialogic Mind: A Dialogic Approach to the Higher Mental Functions", *New Ideas in Psychology,* 1, pp.47-62) and political psychologist Philip E. Tetlock (see Tetlock, 2002, "Social Functionalist Frameworks for Judgment and Choice: Intuitive Politicians, Theologians, and Prosecutors", Psychological *Review,* 109, pp. 451-471) are also reviewed in Bargh's development of his general standpoint. Fernyhough aims to give an account of the higher mental functions characterizing unique aspects of human thought, such as the ability to form concepts, which stresses their essentially "dialogic" nature and the fact that they are grounded on social development. This is his summary of his view:

"[T]he higher mental functions develop through the progressive internalization of semiotically manifested perspectives on reality, such that mature functioning involves the simultaneous coming-into-conflict of differing internalized perspectives. As these perspectives are derived from interaction with actual people with actual positions in the world, they include ontological, axiological, conative and motivational elements. By taking on the voice of the other, the individual also takes on the perspective manifested by that voice, resulting in a form of mental functioning that consists of an ongoing dialogue between differing perspectives on reality". (Fernyhough, 1996, p.53).

Different perspectives on reality also seem to be at the heart of Tetlock's proposed social-functionalist frameworks. Each of these frameworks is meant to serve as "a guiding metaphor that captures the essence of a particular functional orientation that the vast majority of people can, under the right activating conditions, adopt toward the social world".(Tetlock, 2002, p.452). There are five distinct archetypes suggested by Tetlock as characteristic of these different frameworks: the "intuitive scientists", who seek causal explanations of the phenomena observed in their environment that can help them make accurate predictions; the "intuitive economists", who seek to maximize the utility of their actions; the "intuitive politicians", who seek to maintain their identities in their social environment by adhering to certain standards of accountability shared with others; the "intuitive prosecutors", who seek to enforce these standards of accountability by identifying violators of common norms; and finally, the "intuitive theologians", who seek to maintain the validity of common practices depending on shared norms.

Tetlock argues that research has focused almost exclusively on the first two of these perspectives, largely ignoring the latter three. By focusing on the perspectives of the intuitive politician, prosecutor and theologian, Tetlock makes the case that people's judgments and choices, and in more general people's thoughts, feelings and actions that look incoherent from the perspectives of the intuitive economist and scientist, make sense from these latter perspectives. That is because the latter perspectives capture aspects of human thought and action that are not captured under the former. Moreover, different perspectives have different goals and motivations associated with them. This implies that a given action, for example, might express one perspective to the expense of another, with the result that there is a clash between different standpoints on what actions are best to take in a given circumstance. Tetlock identifies a variety of potential conflicts between these perspectives. The intuitive scientist's goals for truth and accuracy might clash with the intuitive theologian's goals to uphold the validity of shared norms, or the intuitive politician's goals to live up to expectations of accountability for his actions might clash with the intuitive economist's goals towards maximizing the expected utility of these actions.

The common theme that runs through Bargh, Fernyhough and Tetlock's standpoints is that our interpersonal development, based on interaction among different perspectives on the world that can be internalized by us and have an effect in our self-understanding and actions, is essential in shaping the way we think and act.

Even though Bargh is not as philosophically oriented as Dennett and Velleman and has played an active role in conducting empirical research into the extent to which unconscious factors influence human behaviour, he is similar to them in that he recognizes that one of the most crucial characteristics of the human mind is that it is the product of both biological and cultural evolution and that it is grounded on a constant interplay among different perspectives, each of which has different cognitive, emotional and behavioural effects associated with it. I think Bargh's standpoint can give us a better handle on explaining the way in which our training as agents influences our actions, even in cases where responses are elicited through coming into contact with features of our environment that trigger unconscious processes that influence our behaviour. Priming of this sort is influenced by our interpersonal development because it is mediated by the conceptual structures that emerge during this social training. Priming the stimulus "generous" to subjects, to go back to his example, would not have the effects it does in their actions if these subjects hadn't developed certain interconnected associations that are activated by this stimulus.

However, a closer comparison between this view and our own account can potentially complicate matters. The main hurdle is that the view that human reasoning consists in a continuous interplay among distinct internalized perspectives might seem incompatible with the idea that self-regulation, construed as fitting one's self-concept to one's actions, is part of what it means to act as an agent. Going back to our own account, the claim is that what makes a behavioural pattern into the action of an agent is that it fits the agent's self-concept which expresses the agent's reasons for acting. But is the answer that clear cut? Perhaps the action examined coheres with a specific perspective, but this perspective is one amongst many. What makes one specific perspective part of an agent's self-concept and not another? Instead, if we claim that all such conceptual structures are part of the agent's self-concept, how is a conflict between internalized perspectives resolved?

My own view is that complex explanations of the nature of the human mind do not need to translate to similar explanations on what it means to act as an agent. As I also discuss in Chapter 3, the way we are cognitively constituted might have as a basis a significantly complex interplay among autonomous processes. Whether we construe these processes as internalized perspectives, conceptual structures or unconscious mechanisms, the point still stands that there is a unity underlying our actions that does not necessarily have to map on any kind of internal simplicity. The argument we borrowed from Pettit can still be applied in this case. The exact nature of the mechanisms leading to our expressions of agency and of reasons for acting is not as important as our capacity to treat our actions as such and our expectation that others will act as such as well. As long as we keep in mind the importance of this capacity, I think that accounts such as the one Bargh provides in his (2006) can help us understand the messier picture of the complex mechanisms underlying our training as agents which enable us to develop this ability to be guided by and provide satisfying reasons for acting.

Furthermore, as I will also argue in the following section on the interplay between conscious and unconscious processes in the context of our account of agency, I do think there's room for determining in more detail cases in which an action expresses an agent's self-concept and when it doesn't, by referring to the content of our self-concepts in given circumstances. Consistently with Pettit's account, I do think that there might be plenty of cases where there are irresolvable vagaries in our attributions of agency, which seems bound to happen during our interactions within a social framework that allows a degree of improvisation in modifying the reasons we provide and in holding ourselves and others responsible for our intentional states and actions. Even allowing these grey areas, I think that there are limits to the extent to which we can do that and there are cases in which certain perspectives, as opposed to others, are more fitting to our self-concept because they better express our judgments on what our circumstances entail.

So where does consciousness fit in?

In this final section, I will place the empirical findings that conscious control and awareness of our actions is limited and that the majority of our behaviour is controlled by unconscious processes under the perspective of our current views on agency. The main idea that we've been led to through exploring the implications of a view according to which agency is a complex phenomenon that emerges from the peculiarities of our social nature is that we should not require conscious acts of will to be the source of every action which is under our authoritative control. Research into what makes a behavioural pattern into the action of a self-knowing agent that starts from the assumption that conscious acts of will are at the source of our actions seems to lead either to rejecting the idea that we can make sense of our attributions of agency to one another or to attempts to argue that, despite appearances, our status as agents is compatible with the empirical evidence. The problem is that there doesn't seem to be a plausible way of choosing the latter option without changing the assumption that drives our questioning, namely that exercises of agency should be identified with direct conscious control of the various processes leading to action. This kind of direct conscious control, as I understand it, would require the agent's being aware of the mechanisms that can lead to a certain action and consciously initiating, or activating, the operation of these mechanisms in order to arrive at the desired outcome. Furthermore, if direct conscious control is all that matters for agency, then the agent seems to be required to consciously monitor and guide the operation of the various processes leading to the course of action he is engaged in.

Having dipped our toes in the empirical literature, we should realize that this view is just not plausible, if one wants to respect the validity of the empirical findings and not simply postulate a desired compatibility between the view involving direct conscious control of action and these findings, without being able to justify this postulation. That's because the prevalence of this kind of hands-on conscious control is not supported by a sophisticated understanding of our nature as empirical creatures whose responses to their environment result from the operation of enormously complex internal machinery. We are only aware of a tiny fraction of these operations, not just because our conscious awareness of them is limited, but because it would not be practical for us, in our continued interaction with our ever-changing circumstances, to have to exercise the kind of arduous conscious control that we can exercise in conscious deliberative reasoning, for example. As our preceding discussion revealed, our success in responding to our environment largely depends on being tuned into it in ways that don't require that kind of conscious control. For example, we can engage in complex skills much more efficiently without direct conscious control and our social interactions can be significantly facilitated by the unconscious operation of processes such as our tendency to imitate one another's behavioural gestures.

Fortunately, as it's become clear though the development of our present account of agency, we don't need direct conscious control to be the centrepiece in our understanding of what makes a pattern of behaviour into an action expressing agentive control. Instead, we need to focus on our social nature and realize that maintaining our own and one another's status as agents is an essential aspect of this nature. We can express authoritative control over our actions because we are guided by a common understanding of how such control is expressed, which depends on being able to engage in an exchange of reasons for acting and to be guided by these reasons. The nature of this understanding has also been elaborated, with the help of the account mainly developed in the preceding chapters. We make certain normative judgments that have the function of reasons for acting. We can regulate our actions in accordance to these judgments that are part of our self-concepts and that express the intentional states that we take ourselves to have. These intentional states can be expressed in a public language and are subject to the norms implicit in our folk-psychological understanding of what it means to coherently express these attitudes in our actions. The ways in which we can regulate our behaviour to express these states are, as such, unique to our social nature.

Besides jettisoning a picture in which direct conscious control is the only thing that matters for our agency, I think we are in a position to nevertheless say more about the role conscious awareness and control might be playing to enable these expressions. I don't think that the account developed so far needs to lead to the conclusion that the aspects of our cognitive structure that are conscious have only an epiphenomenal role. In fact, I think conscious awareness and control are still essential to our status as agents. That's because despite the fact that the main work leading up to our actions is done by the operation of unconscious subpersonal mechanisms, most of which we don't directly consciously control, the way this work is carried out can change depending on conscious aspects of our cognitive structure. As I'll argue, conscious awareness of our behaviour and conscious reasoning, for example, can still influence the way we express ourselves in our actions, though not in as direct a way as the picture according to which we directly consciously initiate this behaviour would have it. Furthermore, I think that our ability to consciously monitor and control some of our behaviour is still essential to our agency, despite its limits. Saying that conscious processing is limited should be taken as a warning against using it as the simple answer to every threat to our ability to act in a self-controlled manner, not as a claim that conscious processing is epiphenomenal.

When asking what makes a behavioural pattern into the action of an agent, we rejected the easy answer of only identifying exercises of direct conscious initiation and control of our actions with exercises of agency. The more sophisticated standpoint that we've developed, mainly with the help of authors such as Pettit, is that we should take into consideration the whole picture and examine the overall framework in which the phenomenon we wish explained takes place. Doing that, we are in a position to recognize that we do have the capacity to act as agents, but that this capacity depends on more than direct exercises of conscious control and we may still frequently fail to express ourselves as agents because our environment and the limits of our empirical nature might interfere with these attempts. When we do act as agents, it is not because we can somehow overcome our empirical nature and control it from a detached perspective, but because we accept our limitations and are able to treat certain behavioural patterns resulting from the interaction between our internal machinery and our surroundings as our own actions.

After rejecting the easy answer and resolving to take the good and the bad involved with being the kinds of creatures we are, what more can we say about the role of conscious processes in our account of agency? We've stressed social influences involving the common understanding of what it means to be an agent, normative judgments and self-regulation that takes the form of fitting one's actions to one's self-concept as key components of this account. What aspects of this account depend on conscious processing? First of all, since the use of a common language depends on conscious awareness of how to communicate publicly with one another, it seems that the contents of our self-concepts that we can express as such should be viewed as something we are consciously aware of. More to the point, we are consciously aware of the intentional states expressed by our judgments contained in our self-concepts, since we can publicly formulate these intentional states. Furthermore, since when we regulate our behaviour to reflect the judgments we make we are influenced by our folk-psychological understanding of how to publicly display the attitudes contained in those judgments, we depend on our conscious awareness of the intentional content of these judgments. It seems right to me to say, as such, that the parts of self-concepts that are conscious consist in the expression of such intentional attitudes that we can verbally communicate and that are governed by norms ingrained in our folk-psychological understanding of intentionality.

Another way to put this is that there is some information contained in our self-concepts that depends on our conscious awareness of how to manifest certain attitudes in our behaviour in an intelligible way. This conscious awareness depends on the manner in which we have been trained to understand agency in our society. But it also depends on the individual attitudes that we express in our judgments. This conscious content is then used by us when regulating our behaviour in order to act in ways consistent with this content. A useful way to see this is that there are certain conscious guidelines that we try to satisfy in order to engage in actions expressing our normative judgments, or reasons for acting. That is because the content of these judgments is conscious, since it can be formulated as communicable intentional attitudes. The agentive behaviour we engage in might still be largely orchestrated by the operation of processes which we do not directly consciously control, but it still depends on these conscious guidelines, since it depends on the use of content which we can consciously formulate.

What about the effortful kind of conscious control, that is associated with processes such as deliberative reasoning and learning to first use a skill? Examples we've used involved someone learning how to manifest a swimming technique or someone arduously abstaining from drinking a fourth glass of wine because of his resolution to stop at three glasses. I think that this kind of control still has a place in our account, as long as we accept that its exercise is limited. For example, we might be able to exercise such effortful conscious monitoring and control of our actions so that they manifest our best understanding of ourselves in cases where we become aware that our behaviour does not manifest this understanding. In most cases, it seems that this kind of conscious control is too costly to use[116] and it seems to be in our best interest to train our behaviour so that this kind of control is not constantly needed.

[116] See, for example, R.F. Baumeister, E. Bratslavsky, M. Muraven and D.M. Tice, 1998, "Ego Depletion: Is the Active Self a Limited Resource", *Journal of Personality and Social Psychology* 74, pp. 1252-1267 and M. Muraven, D. M. Tice and R. F. Baumeister, 1998, "Self-Control as Limited Resource: Regulatory Depletion Patterns", *Journal of Personality and Social* Psychology, 74, pp. 774-789.

To go back to Bargh and Chartrand's point on the beneficial aspects of automaticity, I agree with them that processes such as unconscious goal activation and control can, in a sense, act in our best interests. I previously argued that when these authors made this argument they did not have a clear conception of what counts as one's best interests and of the way an action can be viewed as an agent's own even if it is automatically produced. I think this vagueness is to some extent dispelled through our own exploration of the subject. The agent's own actions, or the actions that express his best interests, are those actions that express the agent's reasons for acting, by being the enactment of the intentional states that are expressed in the judgments that are part of his self-concept. Dispelling this vagueness allows us to explore more effectively cases in which the agent acts besides himself, or besides his best judgment, and cases where he expresses himself through his actions. In terms of conscious content and control, we can say that unconscious processes leading to an action that expresses the agent's self-concept still adhere to certain conscious guidelines, since they still make use of the conscious content in the agent's self-concept. In some occasions, this use might require a kind of effortful conscious control, as in the cases discussed previously, but a lot of the time it seems that this content is used automatically because of the way the unconscious processes have been trained to function.

In a similar spirit, we can make sense of cases in which one's agency is not expressed in one's actions. These cases might be described as ones where the behaviour displayed does not match the agent's normative judgments. There might be a subset of such cases that involve some kind of self-deception, where the agent somehow ignores his normative judgments or overestimates the extent to which he is able to express these judgments in his actions. I will not presently discuss these cases in more detail or attempt to provide a more precise explanation of what it means to act despite oneself. It is sufficient at this point to argue that we can at least make some initial distinction between cases in which agency is expressed and cases in which it isn't, even though the nature of cases of the latter sort is not discussed in detail in this chapter[117].

[117] See the next chapter for an elaboration on what cases in which agents fail to properly exercise their agency in their actions might involve.

These considerations on the role of conscious aspects of our cognitive nature in our ability to express ourselves in our actions might seem, for now, to be just the product of rampant speculation on my part on how to accommodate the exciting findings in empirical studies with a well-rounded account of agency. My view is that there is significant support for these considerations in the current strands of thought in the relevant literature on agency and self-control. In addition to the views already discussed, I think there are other interesting standpoints that can enable us to better appreciate the plausibility and possible implications of our present account. By way of concluding this chapter, I will discuss two such standpoints: John Haidt's (2001) discussion of the interplay between agents' conscious reasoning and their gut feelings when arriving at moral judgments[118] and John Bargh's (2005) latter view on the interaction between conscious and unconscious processing[119].

In his (2001) "The emotional dog and its rational tail: A social intuitionist approach to moral judgment", Haidt sketches a "social intuitionist" model according to which the moral judgments people make are usually being formed on the basis of unconscious intuitions, or gut feelings, and not on the basis of conscious reflection on what the proper response to a moral issue is. According to this model, the judgments reached because of the agent's intuitions come before the conscious reasoning that supports them, even though the agent might feel that it was his own conscious reasoning that led to the judgments being formed. In the context of this account,

> "[m]oral reasoning is usually an ex post facto process used to influence the intuitions (and hence judgments) of other people. In the social intuitionist model, one feels a quick flash of revulsion at the thought of incest and one knows intuitively that something is wrong. Then, when faced with a social demand for a verbal justification, one becomes a lawyer trying to build a case rather than a judge searching for the truth." (Haidt, 2001, p. 814)

This model, as Haidt is quick to clarify, does not imply that conscious reflection on one's judgments or conscious effortful reasoning is useless[120]. Even though most of

[118] See John Haidt, 2001,"The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment", *Psychological Review*, 108, pp. 814-834.

[119] See John Bargh, 2005, "Bypassing the Will: Towards Demystifying the Nonconscious Control of Social Behavior", in R.Hassin, J.Uleman and J. Bargh, (eds.), *The New Unconscious*, New York: Oxford, pp. 37-58.

[120] See, for example, Haidt, 2001, p. 819.

the time, according to him, moral judgments depend on the gut feelings agents get when they are faced with certain situations, conscious reflection on these judgments can be useful since it might produce new intuitions that can lead to different moral judgments. Communication is also useful since it might frame a moral problem in a way that will lead to new intuitions about what the right answer to that problem is. In rare cases, intuitions an agent has can be overridden by the agent's own conscious reasoning and the conclusions it leads to, but these cases as Haidt postulates are rare, restricted mainly to agents who have been trained to accept the results of rigorous reasoning (philosophers being his main example) even when it leads to counter-intuitive conclusions.

Haidt also offers some suggestions towards ways in which agents might manage to increase the influence their conscious reasoning has on the judgments they make. For example, a society whose members train themselves and others to engage in a thorough examination of the various evidence they consider when making judgments about their circumstances, to frequently reflect on the judgments they make and to seek input from other agents on these judgments might be a society which cultivates more rational intuitions in its participants. These intuitions might be more rational in the sense that they are better supported by the available evidence and involve less of a bias in selecting arguments in favour of the judgments formed on the basis of these intuitions. Haidt argues that "[a] more intuitionist approach is to treat moral judgment style as an aspect of culture, and to try to create a culture that fosters a more balanced, reflective, and fair-minded style of judgment." (Haidt, 2001, p. 829).

Although Haidt's discussion is focused on the moral judgments people make, such as the judgment that incest is wrong no matter the circumstances[121], I find his understanding of the interplay between conscious and unconscious processes in the production of these kinds of judgements pertinent to our own discussion. My interest in this view is mainly in that it respects the growing evidence leading towards a better understanding of the various ways in which human behaviour is influenced by unconscious factors, while avoiding a rushed conclusion according to which conscious factors play no significant role in action. Furthermore, we can use Haidt's discussion as an example of the way in which our social development can have a

[121] See Haidt, 2001, p.814.

significant impact on shaping the interplay between conscious and unconscious factors in our responses to our environment (in Haidt's case, on the way in which conscious reflection interacts with unconscious intuitions to produce moral judgments). As he argues, "[t]he social intuitionist model therefore….is not an anti-rationalist model. It is a model about the complex and dynamic ways that intuition, reasoning, and social influences interact to produce moral judgment." (Haidt, 2001, p. 829).

To this line of thought, we should add Bargh's (2005) idea that

> "the purpose of consciousness- why it evolved- may be for the assemblage of complex nonconscious skills… Intriguingly, then, one of the primary objectives of conscious processing may be to eliminate the need for itself in the future by making learned skills as automatic as possible. It would be ironic indeed if, given the current juxtaposition of automatic and conscious mental processes in the field of psychology, the evolved purpose of consciousness turns out to be the creation of ever more complex nonconscious processes." (Bargh, 2005, p. 53).

This take on consciousness' main role is compatible with the view that the way we act is to a large extent the result of processes that we are not consciously aware of and we do not consciously control, and that despite this, conscious awareness and control have a significant role to play in shaping our actions[122]. Perhaps we mostly rely on conscious awareness and control in our early social development, and as we become more adept in expressing our agency in our actions and become more conditioned to exhibit certain patterns of behaviour under specific circumstances, our responses to our environment become less dependent on conscious factors. Just as in Haidt's view we might develop the capacity to form more reflective intuitions, assuming we are trained from an early stage to use conscious reflection and communication in the production of our moral judgments, so it is in our view that our

[122] In this context, it's worth noting G.B Moskowitz and his co-authors' experiments on the extent to which subjects with chronic egalitarian goals were influenced by the automatic activation of certain stereotypes (see G.B. Moskowitz, P.M. Gollwitzer, W. Wasel and B. Schaal, 1999, "Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals", *Journal of Personality and Social Psychology* 77(1), pp. 167-184). In these studies, subjects who were shown to have long-term egalitarian goals (who were, for example, motivated to treat both sexes fairly), were able to inhibit the influence of activated stereotypes on their behaviour, in contrast to subjects who didn't share the same egalitarian goals. This was shown to be the case even in circumstances where the subjects' responses were so fast that they couldn't have exerted conscious control on the stereotypes influencing their actions. I think these experiments can fit our present discussion as an illustration of a case in which subjects that have consciously trained themselves to respond to their circumstances in a certain way (in these cases, by having a long-term commitment to egalitarian goals) can act in that way even in circumstances where they are unable to exert conscious control in their behaviour.

responses to our environment might conform to a greater extent to the reasons we publicly (and hence consciously) exchange for our actions and to the common standards of intelligibility that govern our reason-giving practices, assuming we are trained from an early age to regulate our actions so that they fit these practices. As we've seen, such self-regulation takes the form of regulating our actions so that they manifest the dispositions that fit our judgments in our self-concept. The degree to which conscious awareness and control is needed in this activity might be lessened the more adept we become at expressing ourselves in our actions as authoritative agents.

To sum up, this is how we should view the interplay between conscious and unconscious factors influencing our actions, in the context of our present account. The idea at the heart of this chapter is that conscious control as an initiating act of will is not what determines whether we can authoritatively express ourselves in our actions in the manner of self-regulated agents. Instead, it is our capacity to be guided by our reasons for acting and to offer such reasons to justify our actions that ensures we are able to act as such self-controlled individuals and to be viewed as such. This is compatible with accepting that our actions are often the results of processes of which we are not consciously aware and that we do not consciously control. However, conscious awareness and control are also essential for enabling us to act as agents.

My suggestion for how that might be the case is that there are conscious guidelines we follow when being guided by our reasons for acting, since by regulating our actions in order to fit our reasons for acting we manifest intentional states in them that we are in a position to publicly formulate. We are also in a position to justify our actions by situating them within the framework of our folk-psychological understanding of agency. In learning how to act as competent agents, we are trained in communicating within the folk-psychological framework that provides the basis for the common standards of intelligibility that we are guided by in order to understand our actions as expressing our reasons for acting. Such training also depends on conscious communication based on a public language and it involves our learning to regulate our actions in appropriate ways, and this self-regulation might also depend on our exerting conscious effort in order to make sure that our actions fit our normative judgments on how we should act. Finally, we might rely on consciously regulating our actions more in the early stages of our development, while gradually making less frequent use of such control, falling back on it only in cases where we become aware that our responses do not express our reasons for acting[123].

[123] From the authors whose work has been influential for my discussion in this chapter, I think that Pettit and Velleman in specific would also be sympathetic to the view that conscious control of our actions, even though it might not frequently come into play in shaping these actions, is still essential for our capacity to act as self-regulated agents. In his (2007) discussion of our capacity to act as competent reason-guided agents, Pettit argues that

"[a]lthough unthinking habit shapes what agents do, the discipline of reason will be in virtual control so far as it is ready to be activated and take charge in the event of habit failing to keep the agent in line. In that event, at least in general, the "red lights" will go on and ensure that the agent remains faithful to the perceived demands of reason[.]" (See Pettit, 2007, p. 84).

As for Velleman, his (2009) discussion of an agent's enactment of his conception of crying is particularly relevant to our discussion. According to him,

"[t]here is also an intermediate stage between losing oneself in an activity and consciously putting it into action. Even when letting oneself get carried away by a behavior such as crying, one can retain enough self-awareness to pull up short if the behaviour becomes discordant with one's thoughts. In this third case, one's thoughts and one's behavior proceed in parallel, connected only counterfactually by one's readiness to stop if the two should diverge… This ability to think along with oneself in this way, with thoughts that neither follow nor lead one's behavior, depends on a degree of self-knowledge that can be attained only through long practice in the more deliberate, thought-first mode of action." (Velleman, 2009, pp. 24-25, footnote 16).

For more on Velleman's work on this idea, see Velleman, 2007a, "What Good is a Will", in A. Leist (ed.) *Action in Context*, Berlin/New York: de Gruyter, pp. 193-215, and Velleman, 2007b, "The Way of the Wanton", in K. Atkins and C. MacKenzie (eds.) *Practical Identity and Narrative Agency*, London: Routledge, pp. 169-192.

Chapter 5

Flirting With Incoherency: Self-Deceptive Inauthenticity and Other Agentive Breakdowns

Introduction

In the previous chapters, I have argued for an account of human agency as a complex social phenomenon emerging from the interaction among self-enacting agents that need to make sense of one another's actions as coherent expressions of one another's reasons for acting. Agents are self-enacting because they are able to develop a self-concept that expresses their normative judgments, which in turn express the intentional states that they take themselves to have. Being such agents, the way we regulate our behaviour is uniquely shaped by our common understanding of what it means to have such states and to express them in our actions. When our actions manifest the intentional states that we judge ourselves to have, they express our reasons for acting in the way we do. In this sense, our common understanding of intentionality also entails that we understand what it means to have reasons and to express these reasons in one's actions, and it also engenders various expectations of reason-guided behaviour that we use in order to make sense of the extent in which our own and others' actions constitute expressions of agency.

A key concept in this account is self-knowledge, or self-understanding, which is displayed by us as competent reason-guided agents when we manifest our capacity to fit our actions to the way we see ourselves. Self-knowledge, in our story, does not depend on introspective prowess on our part. We do not know our own mind because we can accurately perceive the intentional states that make up our psychological constitution, but because we have the capacity to express the states that we judge ourselves to have in our actions. The self-concept that expresses our intentional states, or reasons for acting, develops through our judgments on what our circumstances entail. Our beliefs, for example, are expressed in our judgments of what is true for us, and our intentions are expressed through our judgments on what is intelligible for us to do in a certain context, given the way we see ourselves as situated in this context. These judgments have a normative status for us because they can be expressed in public claims in a common language. We have been trained, through our upbringing in a social collaborative framework, to make these claims using terms understood by others as expressing our state of mind by referring to states such as hopes, fears, beliefs and desires. In making these claims, we commit ourselves to certain demands of rationality, which in this account are demands to live up to our normative judgments. Expressing our agency in our actions entails having the capacity to justify them as manifestations of our reasons for acting. Being able to competently offer such reasons also entails that we are able to regulate our actions in accordance to these reasons.

An implication of this account is that whether we live up to our self-understanding is something that others can also have a say in. If there are shared standards of intelligibility which guide human action, consistently violating these norms is something that we can be criticized for. If we are found to consistently provide an ill fit between our words and deeds, then others might evaluate this behaviour as failing to express our authoritative agency and challenge the reasons we provide for acting in the ways we do. Critical scrutiny of one another's reason-guided behaviour is thus something all competent agents can take part in and expect each other to be able to participate competently in. Evaluating the degree to which an agent's actions are borne out of genuine self-understanding involves examining the agent's reasons for acting and the degree to which they provide an acceptable justification of his actions. This critical examination is something we can also engage in with regards to ourselves and it seems that we are in a position to recognize not only when others fail to coherently express their agency in their actions but when the same is true of us as well. If we are able to arrive at this recognition, it seems that we should also be able to take steps to do something about our failings, by taking steps to counter such failings. Our recognition that there is some sort of flaw in our expressions of agency might be facilitated by others' critical interpretation of our actions, and taking steps to accommodate this flaw might also be something that we can do by ourselves or with the help of others.

These initial considerations breed a host of new questions. What is a failure of agency, in this account? Are we talking about just one kind of failure or is there a variety of ways we can fail to act as agents? How would an agent be able to recognize that there is a flaw in the way he expresses himself in his actions? How much is idiosyncratic about this recognition and how much depends on common standards of intelligibility? What kind of steps can the agent take in order to accommodate the flaws in his expressions of agency? What kind of steps can others take in order to help the agent accommodate such flaws? Is there a sense in which some methods of accommodating these flaws would disrupt the agent's capacity to express himself in his actions? Finally, is there a line we can draw between helpful guidance and intrusive manipulation when examining the steps taken to counter failings of agency?

I think that we can make some headway in answering these questions if we apply them to our current account of agency, and more specifically to our current understanding of self-knowledge (see especially Chapter 2). I intend to argue that an important failure of agency consists in self-deceptive inauthenticity and that we should understand failure to act as an agent as a failing in one's capacity to express a coherent self-image in one's actions. Self-deception should be understood as a failure of self-knowledge in that one takes one's self to be doing or thinking something other than one actually does, by falling victim to inauthenticity. The extent to which one expresses a genuine self-concept in one's actions also depends on the shared standards of intelligibility guiding our actions. J. David Velleman seems to hold this view on what inauthenticity amounts to:

> "…[I]nauthenticity involves acting on a false self-conception- a self-conception that one does not succeed in making true by acting on it. Although I think that every action of this kind is inauthentic, the term carries normative connotations that may not be appropriate in all cases. For this reason, we tend to reserve the term for cases in which the false self-conception is adopted self-deceptively, in order to avoid some unpleasant truth about oneself." (Velleman, 2009, pp.60-61, footnote 2)

It's interesting to compare this view with Sartre's "bad faith."[124] For Sartre, inauthenticity, or bad faith, seems to consist in behaviour that does not express the

[124] See Stephen Priest (ed.) *Jean-Paul Sartre: Basic Writings,* Routledge, especially pp. 204-220.

agent’s self-determined freedom to make the choices he does. From Sartre’s examples, it seems that this can occur if the agent has a flawed understanding of himself and the context in which he finds himself. Treating one’s self as a determined object that has no choice but to act in the ways it does is one of the ways in which Sartre argues that the agent’s behaviour can be understood as an expression of inauthenticity. It seems that in his view, even adopting a social role for the sake of others that constrains one’s actions expresses a flawed self-understanding and leads to inauthentic behaviour. In discussing Sartre’s examples (the woman who dissociates herself from her body to avoid unwanted attention, the man who plays the role of a waiter), I’ll agree with Sartre that seeing one’s self as completely driven by external determining forces leads to self-deceptive inauthenticity, but disagree that all cases in which the agent adopts a social role for the sake of others count as cases of bad faith. In discussing these cases I will adopt Velleman’s understanding of Sartre and his argument that enacting a role is not inauthentic in so far as the agent is aware of himself as the enactor of this role and does not treat this role as something distinct from himself that drives his actions[125].

Another aspect of inauthenticity that I think Sartre’s discussion of bad faith can help us delve into concerns his distinction between two perspectives on one’s self, that of “facticity” and that of “transcendence”. Facticity involves seeing one’s self as an empirical object and realizing that one is constrained by the same empirical factors that constrain other such objects. Transcendence involves seeing one’s self as a decision-maker that is able to express his freedom of choice in his actions. Sartre argues that inauthenticity can result from taking the one perspective for the other and acting as if one is motivated by considerations supported through the use of one perspective in order to avoid facing the consequences of adopting the other. I have discussed these two different perspectives in Chapter 2 in comparing the views of Richard Moran[126] and Victoria McGeer[127] on self-knowledge. I think these authors’ work can be illuminating also in this context, as their views can help with understanding how self-deceptive inauthenticity can arise from the interplay of these two perspectives. In discussing these views, we will be in a position to see that self-

[125] See Velleman, 2009, especially pp.25-26.
[126] See Moran 1997, 1999-2000, 2001.
[127] See McGeer 1996, 2007a, 2007b.

deceptive inauthenticity can result from either treating one's self as a determined object that will run its course regardless of one's choices or from acting as if one's choices are entirely unconstrained by the kind of creature one is. I think the latter cases of self-deception are especially interesting, since they include what one might call rampant rationalization. McGeer offers an example of such a case in which a self-deceived individual endorses the motives leading to his actions as much nobler than the same kinds of motives found in others because of his flawed understanding of himself and his circumstances.

To conclude, I wish to discuss how the shared standards of intelligibility that govern our actions impact on our view of what it means for an agent to express a flawed self-understanding in his actions. In our account, the degree to which the agent's understanding of himself and his circumstances is determined to be lacking also depends on how efficiently the agent can convey this understanding in his actions in the context of the reason-guided practices that all competent agents participate in. Critical scrutiny of one another's actions makes sense in this normative context, since it allows for the possibility of countering agentive failings such as self-deceptive inauthenticity by bringing self-deceived agents in a position to have a more genuine self-understanding expressed in their actions.

Self-deceptive inauthenticity

Acting in an inauthentic manner can be taken to mean that one acts as something that one isn't. One way to understand this claim is that one's actions do not express one's "real" self. But this already sounds quite puzzling. We'd have to have a good definition of what a "real" self is before we attempt to describe cases of inauthenticity. In my own discussion, I have avoided views in which the self is some kind of concrete entity in an agent's body that exercises deliberative control over the agent's actions. Instead, my preference (argued for in previous chapters) is for a view in which control is distributed and the closest thing to a self an agent has is the agent's self-concept that expresses the agent's reasons for acting. The self-concept itself does not fit what might traditionally be construed as the self, since it doesn't itself exercise any kind of authoritative control over the agent's actions. What makes such a theoretical construct interesting is that it is useful for the agent as a guide to his own reasons for acting, and so it is used by the agent in regulating his actions so that they express these reasons. The self-concept is the closest thing to a self the agent has because actions guided by the information contained in it express the agent's own point of view, because they express the agent's judgments on his circumstances. Considerations of infinite regress do not threaten this account, (as they would, were the nature of the agent who uses the self-concept as mysterious as the nature of the traditionally construed self), because the self-concept is used by distributed information processes making up the agent's cognitive organization.

According to this view, for the agent to act as something that he's not involves a failure in the way the agent uses his self-concept, since expressing one's self in one's actions involves acting in accordance to one's self-concept. By acting in accordance with his self-concept, the agent acts in accordance with the judgments that express his reasons for acting. So failing to act in accordance to one's self-concept involves failing to express one's reasons for acting in some way. This is a failure of self-understanding on the agent's part, since he cannot be said to know his own mind if the reasons that he takes to be motivating his actions are not in fact expressed in his actions. But how can the agent be mistaken in this way? Does the agent have a genuine self-concept that he is unable to act in accordance with, or does he act in accordance with a self-concept that does not express his genuine reasons for acting?

Cases of the first sort, in which the agent has a genuine self-concept that is not guiding his actions, are failures of agency that seem to consist in an inability to manifest one's dispositions in one's actions so that they fit one's normative judgments. It seems that we can provide a pretty straightforward explanation for these cases by treating them as examples of weakness of the will. The agent is unable to manifest the dispositions that are expressed by his normative judgments because he is unable to resist the force of motives opposing these judgments, even though he is aware of what dispositions he ought to manifest in order to live up to the self-understanding that best fits his circumstances. I would argue that cases of weakness of will reflect a flawed self-understanding on the agent's part only inasmuch as the agent misunderstands the force of the motives he identifies with. These motives, as our account has it, can be viewed as the agent's reasons for acting because the agent is able to make his actions intelligible by reference to them. In cases of weakness of the will, the agent is unable to exercise his capacity to be guided by his own reasons because he is unable to manifest the dispositions that would make his actions intelligible. The agentive failure in this case is owed more to limits in the agent's capacity to overcome motives that do not express the attitudes that he takes himself to have and less to a flaw in the agent's self-understanding.

Cases of the latter sort, however, in which the agent takes his behaviour to be guided by reasons that he doesn't actually have seem far more puzzling. This kind of failure in the exercise of agency does not seem to adhere to the same explanation as failure due to weakness of the will. In weakness of the will, the agent seems to be aware of his inability to express his self-understanding in his actions and hence he is aware of the fact that his behaviour doesn't express his own reasons for acting. But in cases where the agent is acting inauthentically, he seems to be deceiving himself into treating his behaviour as the result of genuine self-understanding, when in fact that's not the case. We might refer to this kind of failure of agency as resulting from the agent's self-deceptive inauthenticity. The main problem here is that, in the context of our present understanding of what it means to act as an agent, it's not even clear that we can coherently describe the kind of breakdown in one's capacity to act as an agent that is characteristic of self-deceptive inauthenticity. So far, I seem to be constantly flirting with incoherency in my attempts to understand what's behind this agentive failing, and the descriptions I've resorted to seem vague or self-refuting. How can the agent attempt to express a self-concept in his behaviour that does not constitute genuine self-understanding on his part? How can he fail to properly exercise his capacity to act as an agent who is guided by certain reasons, if he treats his actions as manifesting the attitudes that he takes himself to have and hence as being guided by his own reasons for acting?

Given that research on self-deception and inauthenticity has frequently focused on the seemingly paradoxical nature of these topics, it's not really surprising that our account also struggles with providing a coherent explanation of self-deceptive inauthenticity[128]. Sartre is one of the authors looking into these topics and his

[128] An important paradox associated with self-deception that I do not explicitly discuss in the main text arises from attempts to account for self-deception by arguing that there is a part of the agent that intentionally deceives another part, or a self that deceives and a self that is deceived. But how is it that an agent can intentionally deceive himself, by hiding some truth from himself for example, while at the same time being unaware that this deception is taking place? How can the agent be the deceiver and the victim of deception at the same time?

discussion of bad faith, with its associated examples, is particularly relevant. Bad faith is a form of inauthenticity, or pretence, which Sartre refers to as a "lie to oneself" (Priest, 2001, p.208) which consists in "hiding a displeasing truth or presenting as truth a pleasing untruth" (ibid), and illustrates this kind of behaviour by using the examples of various individuals misapprehending their nature. Two of the main cases used by Sartre for this purpose are that of a woman pretending that her date's flirtations are directed at her body, which she takes to be distinct from her, and that of a man pretending that his behaviour is wholly determined by his role as a waiter. Both these individuals are taken by Sartre to exemplify some kind of inauthenticity by concealing some facts about their nature from themselves and, as such, by deceiving themselves, through it's not immediately obvious what these concealed facts are.

To begin with, I will focus on the example of the waiter, as it seems to lend itself to the most unambiguous interpretation of what self-deceptive inauthenticity amounts to. Sartre talks of the waiter as someone who "applies himself to chaining his movements as if they were mechanisms, the one regulating the other" (ibid, pp.218-

Sartre (see Priest, 2001, pp. 208-214) argues against resolving this paradox by using a distinction between an unconscious deceiver and a conscious victim of deception, as one interpretation of psychopathological cases would have it. Sartre argues that the paradox still remains in this case, because it seems that if one's flawed understanding of one's self in one's circumstances is unconscious and consciously repressed, there is still a problem explaining how the conscious part is able to repress the threatening information if it is unable to recognize it as such. It seems that there is still a sense in which the agent as a whole knows that he is deceiving himself, which brings us back to our paradox.

Alfred Mele's (see Mele, 1997, "Real Self-Deception", *Behavioral and Brain Sciences*, 20, pp. 91-102 and Mele, 2001, *Self-Deception Unmasked*, Princeton University Press) views on self-deception are useful for dispelling this paradox, since he argues we should not apply the same understanding we have of interpersonal deception to the case of self-deception, seeing that self-deception can be motivated but not intentional. Furthermore, Mele argues that there is no need to postulate unconscious intentions to deceive one's self when his own theory can explain the same self-deceptive phenomena in a less complicated manner. Another theorist that distances himself from treating self-deception as intentional is Richard Holton (see Holton, 2001, "What is the Role of the Self in Self-Deception", *Proceedings of the Aristotelian Society,* New Series 101, pp.53-69), who argues that self-deception should be treated as a mistake about the self rather than as intentionally deceiving one's self.

In general, I think that approaches such as Mele and Holton's to this paradox are on the right track and that we should steer clear of an understanding of self-deception that treats it as involving a part of the agent that intentionally deceives another part, or a part of the agent that is somehow aware of some fact that it knowingly conceals, and a part that is unaware of the concealment taking place. As I will argue in the main text, cases of self-deceptive inauthenticity should be viewed as involving a flaw in the agent's self-knowledge that leads to self-deceptive behaviour. But this self-deceptive behaviour should not be understood as behaviour intentionally engaged in as deceptive behaviour. Despite any turns of phrase that might seem to suggest otherwise, at no point should my description of what I take to be cases of self-deceptive inauthenticity be taken to imply that I favour an approach that treats self-deception as involving this kind of intentional activity.

219) and who is "playing *at being* a waiter in a café." (ibid, p.219). At first, what lies at the core of the waiter's inauthenticity seems to be the adoption of a social role. In talking of himself as adopting such a role, Sartre identifies the following behavioural predicament:

> "[Being a waiter] is a 'representation' for others and for myself, which means that I can be he only in *representation.* But if I represent myself as him, I am not he; I am separated from him as the object from the subject…I can only play *at being* him; that is, imagine to myself that I am he.....In vain do I fulfill the functions of a café waiter. I can be he only in the neutralized mode, as the actor is Hamlet, by mechanically making the *typical gestures* of my state and by aiming at myself as an imaginary café waiter through those gestures[.] What I attempt to realize is a being-in-itself of the café waiter, as if it were not just in my power to confer their value and their urgency upon my duties and the rights of my position, as if it were not my free choice to get up each morning at five o'clock or to remain in bed, even though it meant getting fired." (ibid, pp.219-220)

As Sartre seems to be saying here, the problem with playing the role of a waiter (or any other role) is that the role robs the agent from his capacity to exercise his freedom over his choices, because the agent is acting mechanically in accordance to the dictates of this role. The man in the example is playing at being a waiter by fitting his behaviour to the role he enacts through treating his movements as mechanisms that are chained to one another and to their initial cause, which is the role being enacted. The agent then is unable to exercise his agency because of adopting a role that does not express his freedom to choose. This interpretation is also shared by Stephen Priest, the editor of "Jean-Paul Sartre: Basic Writings":

> "The reality of our freedom is so unbearable that we refuse to face it. Instead of realising our identities as free conscious subjects we pretend to ourselves that we are mechanistic, determined objects. Refusing to freely make ourselves what we are, we masquerade as fixed essences by the adoption of hypocritical social roles and inert value systems". (ibid, p. 204).

The adoption of such "hypocritical social roles", or "representations", might then be one way in which we can understand an agent's failing to exercise his capacity to exercise his agency in his actions and falling into self-deceptive inauthenticity. The basic fact about their nature that agents are concealing from themselves, under this interpretation, is that they are not determined by the roles they adopt because they are free to make their own choices. In Sartre's view, these roles seem to include any kind of social role or representation that the agent adopts. These roles are adopted for the sake of others and do not express the agent himself as the person choosing and in control of his actions.

Taken as it is, this interpretation seems at odds with our own understanding of agency. Recall that in our account, we all adopt the role of the agent in our interactions, not only for the sake of others but also for our own sake, in order to make our actions intelligible to ourselves. As discussed in previous chapters, we can use J. David Velleman's analogy with improvisational actors in order to illustrate what it means to act as an agent (see especially chapter 3). We all improvise ourselves by consistently enacting the way we understand ourselves in our behaviour. The enactment of our role as agents has to be coherent enough for our behaviour to make sense as manifesting the attitudes expressed by the judgments contained in our self-concepts. Being able to enact the role of the agent in this way, we have the capacity to exchange reasons for acting and to be guided by these reasons, because we have the capacity to treat our actions as intelligible enactments of our role as agents. But if we stick to this view, we are not able to use the aforementioned interpretation of self-deceptive inauthenticity, because we would be led to a contradiction. On the one hand, one way I've argued that we can understand our capacity to act as agents is to see it as the enactment of a social role, namely that of an agent who is able to coherently express himself in his actions by fitting his behaviour to his self-concept. On the other hand, according to one interpretation of Sartre's understanding of self-deceptive inauthenticity, the agent acts inauthentically because of adopting a social role and letting that role guide his actions. To integrate these viewpoints, it seems I'd have to argue that the very thing that constitutes our capacity to act as agents robs us of our agency, so by enacting the role of the agent we fail to exercise our capacity to act as agents. But this doesn't make any sense.

There is a way out of this dilemma though, which involves a different understanding of Sartre's example of the waiter and of the reasons the waiter falls into self-deceptive inauthenticity. I owe this different interpretation of the example to Velleman and his reconstruction of Sartre's discussion[129]. The problem with the behaviour of the man who plays the role of the waiter, Velleman argues, is not

[129] See Velleman, 2009, pp.25-26.

simply that he is enacting a social role. The problem is in the specific way in which that man plays the role of the waiter. He is enacting a role that doesn't express him, as an agent, and he treats this role as something separate from himself that determines his behaviour. Velleman's point here is that the man underestimates his capacity to act as an agent, because he fails to treat the role of the waiter as his own role, performed willingly as the role of a self-enactor, and in so doing falls into self-deceptive inauthenticity. The main fact this man conceals from himself is that he is not merely a series of interlinked mechanisms, driven by the waiter's role, but instead these mechanisms are put into motion because of his own enactment of the role of the waiter. According to Velleman's reconstruction of the example,

> "the waiter.. plays the role of a waiter as if it weren't a role, as if his inherent waiterliness were directly controlling his movements, whereas he is actually conforming those movements to his conception of the waiterly thing to do. The waiter would not be in bad faith if he let go of self-awareness and fell back on his professional habits and skills, proceeding on 'automatic pilot', or if he enacted the part of a waiter candidly, by playing a self-enacting waiter who is admittedly fitting his behavior to a conception of what a waiter would do. What lands him in bad faith is that he plays the part of a waiter who isn't playing the part." (Velleman, 2009, pp. 25-26)

These two interpretations are similar in that according to both, the man playing the waiter is mistaken about the nature of his behaviour when he treats it as a mechanical display that will run its course regardless of his own contribution. The difference is that, while Sartre seems to identify the source of this misapprehension in the adoption of social roles in general, Velleman finds it only in the adoption of roles that the agent treats as distinct from his own agency. In the context of our own account, the latter interpretation is clearly preferable, because it helps us take a step towards understanding inauthenticity without descending into either incoherency or having to discard some of the key ideas structuring our previous discussion. At the same time, choosing this interpretation can also enable us to preserve the importance of the similarity of these viewpoints and draw the following conclusion: Self-deceptive inauthenticity does not involve the adoption of social roles in general, but it seems to involve a flaw in the agent's self-understanding. One of the ways in which the agent can make a mistake of this sort, which both Sartre and Velleman emphasize, is by treating his behaviour as wholly determined by something other than his own agency.

That this "something other" happens to be the role of a waiter in Sartre's example is coincidental to the fact that the self-deceived agent dissociates himself from his actions by treating them as the product of independent mechanisms that don't require his own contribution to produce their effects. This flawed self-understanding might not have resulted from the man's adoption of a social role. This waiter might have instead seen his actions as the direct result of his boss' orders, which he takes to be the main driving force behind his every move. These orders, according to the latter understanding the waiter would have of his situation, directly feed into his actions independently of any choice on his part on how to proceed. This doesn't involve this man's taking his role as a waiter as the main determinant of his actions, but it would still entail that he ends up understanding his behaviour as that of a determined object that does its thing regardless of his choices on the matter. In this case, the waiter's understanding of his nature would still be flawed and lead to a failure on his part to properly exercise his capacity to act as an agent. That's because the waiter would fail to understand that his behaviour is not the direct result of his boss' orders independently of his choice to follow these orders, in the same way that it's not wholly determined by his role as a waiter independently of his choice to perform the duties of a waiter[130].

The other main example used by Sartre as an illustration of bad faith also seems to fit this pattern. This is the case of a woman who refuses to recognize that the man she is speaking to might have more than a platonic interest in her, and ignores any implications of his actions that would threaten her understanding of her situation. When he takes hold of her hand, instead of recognizing this action as an expression of attraction on the man's part, which would mean she would have to make the choice of how to respond to his touch, she just treats her hand as out of her control.

[130] See also Velleman, 2009, pp.25-26, footnote 17, on identifying the main source of the flaw in the waiter's self-understanding in his construal of his behaviour as that of a determined object that is independent from his own agency:

"I think that Sartre is less than clear about the nature of the waiter's bad faith. Sartre says that the waiter is in bad faith simply in virtue of 'playing at being a waiter'; but he also points to the deliberately mechanical style of the waiter's movements as symptomatic of his bad faith. As I see it, this simulated automaticity shows, not that the man is playing at being a waiter, but rather that he is playing at being a waiting-machine- that is, something that does what a waiter does but without enacting an idea of it."

As Sartre describes the case, by refusing to respond to her date's sexual interest, she treats her body "as a passive object to which events can *happen* but which can neither provoke them nor avoid them because all its possibilities are outside of it." (Priest, 2001, p.215). This woman also seems to fall prey to the same flawed self-understanding as the self-deceived waiter. By treating her body as an object impervious to her choices, she fails to see that it's up to her to leave her hand in the man's hand and in so doing, fails to exercise her agency in her actions.

Both these cases then fit the interpretation according to which treating our behaviour as that of a determined object that is not influenced by our own decisions and choices constitutes a flawed understanding of our nature and ultimately leads to self-deceptive inauthenticity. How can we place this conclusion more explicitly in the context of our own account? I'd argue that an agent forms a flawed self-concept when he treats his intentional attitudes as occurring independently of his own judgments on what his circumstances entail. In other words, the agent is in a position to make his attitudes explicit in his actions because of the way he judges he should act. His actions express his reasons for acting because of his capacity to fit his actions to his self-concept, by manifesting the attitudes expressed in his judgments. What seems essential to the agent's having a genuine self-understanding is his realization that he would not have the attitudes he does if he wasn't able to manifest them in his actions so that they fit his normative judgments and so that they make his actions intelligible as expressions of his reasons for acting. It seems that one way in which the agent acts inauthentically is when he does not take into account the fact that his actions depend on the reasons he has for acting, and that if he didn't have these reasons he wouldn't act as he does.

I think we can make more sense of this interpretation of Sartre's cases, and expand our understanding of self-deceptive inauthenticity, by considering Sartre's distinction between the two different perspectives on human nature he calls "facticity" and "transcendence". As we'll see, the main agentive failing we've identified in the aforementioned cases of bad faith can be understood in the context of these perspectives, and is not the sole potential source of an agent's self-deceptive understanding of his nature.

Facticity and transcendence as sources of self-deception

In discussing the case of the self-deceived woman who is the target of her friend's unwanted attention, Sartre also provides the following observation:

> "We have seen also the use which our young lady made of our being-in-the-midst-of-the-world- i.e., of our inert presence as a passive object among other objects- in order to relieve herself suddenly from the functions of her-being-in-the-world- that is, from the being which causes there to be a world by projecting itself beyond the world toward it own possibilities." (Priest, 2001, p.217)

With this remark Sartre refers to the two perspectives directed at what he describes as "the double property of the human being, who is at once a *facticity* and a *transcendence*." (ibid, p.215). From the perspective of facticity, the agent sees himself as an empirical object that is similar to other such objects and is affected in the same ways that they are. The agent is also in a position to transcend this empirical nature, by realizing that the behaviour exhibited by what is identified, from the perspective of facticity, as an empirically constrained object, also depends on his choices and decisions. The perspective of transcendence is directed at this self-determined aspect of the agent's nature and treats the agent as the source of his actions, as the one who makes decisions and chooses which courses of action to pursue. Both these perspectives are valid as they correspond to different aspects of an agent's nature, but they can potentially enter into conflict with one another.

These different perspectives on human nature are also central to Richard Moran's work on the nature of agency and self-knowledge, which I've discussed in a different context in Chapter 2. Moran borrows these perspectives from Sartre and shares the view that these are two different stances than an agent can adopt toward himself. He frequently refers to these as the "theoretical", or "empirical" stance, and the "deliberative" stance, which correspond to Sartre's perspectives of facticity and transcendence, or the agent's first-person and third-person perspective on his actions, respectively. The deliberative stance is the stance from which the agent forms his attitudes by focusing on the reasons he has for them and the one from which he makes his decisions to act. The agent can also view himself from a third-person standpoint, through which he is in a position to recognize his empirical limitations and his objective similarities with other agents who are likewise empirically constrained. The use of the empirical stance does not depend on the agent's subjective, first-person understanding of his mind and actions, since external observers of his behaviour can also use the same stance in order to attribute various mental states to him and interpret and evaluate his actions.

Motivated by these considerations, Moran argues that the agent doesn't gain, through the use of the empirical stance, any kind of privileged insight into his mind and actions that others will always lack. Any such insight the agent gains is always, in principle, available to external observers as well, who can also examine the same kinds of data that the agent examines in order to arrive at various conclusions on the nature of his actions and mental states. Part of Moran's point is that the agent's conclusions are no more privileged because of his capacity to introspect on his mental states. Both introspection and external perception of the agent's behaviour are based on a theoretical standpoint, from which the agent's psychology and its effects on his actions can be evaluated as objective facts that are available to anyone in a position to examine the agent's behaviour and his psychological constitution.

Even so, the agent can still have a different kind of self-knowledge and control over his thoughts and actions, without relying on any kind of privileged capacity to read his own mind. The agent can act as the self-knowing, self-controlled author of his own actions, as Moran's argument goes, because he is in a position to endorse certain reasons and express these reasons in his actions. The agent can only do that through engaging the deliberative stance and focusing on the reasons he has for his attitudes and actions, instead of treating these actions and attitudes as events occurring independently from his own deliberation on his circumstances and on the courses of action dictated by these circumstances. This puts the agent in a special relation to his own mental states and actions that isn't shared by anyone looking at his behaviour from a third-person standpoint. This special relation comes from the fact that the agent's psychological constitution and actions can reflect his deliberative conclusions.

For example, the agent's coming to a deliberative conclusion on what is true in a given situation leads to a corresponding belief being formed that expresses the agent's resolution. The fact that the agent's psychological constitution includes this belief, Moran would argue, is due to the agent's making up his mind on the matter. The main difference between third-personal observation and first-personal deliberation, for Moran, is that the observation treats what is being observed as an objective fact that is independent of the observation itself, while the results of the deliberation are constitutive of what is being observed from the third-personal standpoint. The agent, by adopting an objective empirical standpoint on his psychological constitution and actions, treats them as something separate from himself, while by adopting a first-personal standpoint he is endorsing them as part of who he is because he is expressing his agency over them.

The deliberative stance, as described, is clearly essential to acting as an agent. For both Sartre and Moran, the agent must recognize that it's up to him to revoke or maintain his attitudes and endorse his actions, by maintaining or revoking the reasons he has for thinking and acting in the ways he does. This, however, as both authors seem to recognize, does not entail that the empirical stance is inessential to one's agency. Both stances are valid in their own way, even though they might come into conflict. If acting as an agent involves learning to both recognize one's empirical limitations and to recognize that one's psychological constitution and actions also depend on the exercise of one's agency, then acting as an agent involves both the empirical and the deliberative standpoint. Conversely, failing to act as an agent involves a failure to maintain some kind of balance between these two standpoints, so the ways in which these stances can come into conflict for the agent are particularly relevant to our current discussion. Through examining this conflict, it'll become clear that acting inauthentically involves the agent's adopting one of these distinct standpoints to the exclusion of the other, in a way that is detrimental both to the development of genuine self-understanding on his part and to his capacity to exercise his agency in his actions.

Going back to Moran, despite the fact that his main focus is on drawing attention to the importance of the deliberative stance, he is also concerned with showing that both stances are valid in their own way and that the agent must maintain a balance between the demands of both stances in order not to undermine his own agency. The following passage, for example, illustrates this concern:

> "[E]ach perspective presents its own demands as unavoidable, requiring an answer in its specific terms. On the one side, the Theoretical perspective tells

> [the agent] to be empirically realistic about himself, and that anything less than this can only be an attempt to make a virtue of his capacity for pretense or wishful thinking. But for all that, it cannot tell him when such “realism” is simply the appearance taken by his acquiescence, or his avoidance of the practical question before him. On the other side, the Deliberative perspective tells him that he is not bound by his empirical history, that he must answer the question of what he will do as a question of what he *is to do,* and that anything less than this can only be a form of evasion. But at the same time, this perspective cannot tell him when his assumption of agency is a mere sham-when, for empirical reasons, he has lost the right to form an intention with respect to this question and expect that to count for anything. Neither perspective *denies* the truths of the other. The assertion from the Deliberative stance that “I am not *bound by* my empirical history” is not in any way a denial that the facts of my history are what they are. It does not deny either the truth of these claims or their relevance to the question at hand; but it does deny their completeness and, in a word, their decisiveness.” (Moran, 2001, p.163)

According to Moran’s understanding of the conflict between the agent’s distinct standpoints on himself, it seems that on the one hand, the agent can use his empirical nature as an excuse in order to avoid making a decision on how to act, while on the other hand, an agent might delude himself into thinking that only his deliberation matters for how he acts in a given situation, regardless of his limitations. Both the deliberative and the empirical stance can lead to a flawed self-understanding, if the agent resorts to adopting one of these stances in order to evade the truths evident in the other.

Keeping Moran’s distinction in mind, it’s easy to identify at least one side of this conflict in Sartre’s example of self-deceptive inauthenticity. The man playing the role of the waiter neglects the duties and responsibilities inherent in the role he plays because he treats this role as a disposition occurring in his behaviour regardless of his own deliberation as a waiter. The flaw in his reasoning is his oversight of the fact that his behaviour is still a result of his own deliberation on what is expected of him as a waiter. He just fails to take this fact into account and realize that his disposition to respond as a waiter when serving his customers is up to him, because he is the one who makes the choice to perform the duties of a waiter. His role as a waiter would affect his actions whether he understood these actions as expressing his own deliberative conclusions or not. The difference is that when acting under bad faith, the man misapprehends the source of this role’s affective power, by mistakenly attributing it to the role itself and not to his deliberative conclusions. In doing so, he seems to think of his disposition to act as a waiter, which is part of his psychological constitution, as an event that is not caused by his own deliberation on how to act and that has its effects in his actions regardless of his own endorsement of these effects.

The form of evasion that Moran links to adopting the empirical stance in place of the deliberative stance is exemplified in this kind of behaviour. By attributing his behaviour solely to his role as waiter, and refusing to acknowledge that his disposition to act as waiter depends on his deliberation on how to respond to his circumstances according to his duties as a waiter and not to any effect this role might have independently of this deliberation, the man evades any responsibility for performing these duties. A similar form of evasion can be identified in our other case of bad faith. By acting as if leaving her hand in her date's hand does not depend on her own deliberative endorsement of the sexual intent in her friend's behaviour, the woman evades the responsibility of making a choice on the matter and mistakenly treats her hand as an object uninfluenced by any decisions she might make.

But what about the opposite kind of evasion? On the one side of the conflict between the empirical and deliberative stance, the agent avoids taking responsibility for his actions and fails to acknowledge that his psychological constitution is also shaped by his own deliberative conclusions on what his circumstances entail. On the other side, the agent might fail to acknowledge his own empirical limitations and he might consider his deliberative conclusions as the only determinant of his actions. Moran seems to understand this kind of agentive failing as a misapprehension on the part of the agent of the extent to which his empirical history affects his capacity to make his deliberative conclusions play a role in his actions. This is what Moran talks about as the agent's "attempt to make a virtue of his capacity for pretense or wishful thinking." (ibid). Cases of this sort seem to be closer to weakness of the will than other kinds of agentive failings, with the main difference being that in weakness of the will, the agent realizes that he cannot act the way he judges he should act because of persistent flaws in his psychological constitution, while in these cases the agent ignores such flaws and pretends that he is unconstrained by them. Think of the difference between two compulsive smokers, the one frustrated by his inability to quit smoking even though he judges that he should, and the other continuing to smoke while claiming that he can quit at any time. The former is weak-willed, while the latter pretends that his deliberative conclusion on whether he should smoke can have an immediate effect on his behaviour despite his compulsive disposition to smoke.

Deluded determinism and rampant rationalization

By this stage, we have identified three kinds of agentive failings. The first is weakness of the will, in which the agent fails to act the way he judges he should and fails to express his self-understanding in his actions because of motives opposing his normative judgments. The agent deliberates on what his circumstances entail but his deliberative conclusions are frustrated by the stronger motives in his psychological constitution. In these cases, it seems plausible to claim that the agent is aware of his inability to properly express himself in his actions because he notices the discrepancy between the way he behaves and his normative judgments that express the intentional states that would make his actions intelligible for him and for observers of his behaviour. As these normative judgments, or deliberative conclusions, express the agent's reasons for acting, we can say that in cases of weakness of the will the agent is unable to make his actions fit his reasons for acting and is aware of his inability to do so.

The second and third kind of agentive failings can fall under self-deceptive inauthenticity, which differs from weakness of the will in that the agent deceives himself into thinking that he expresses a genuine self-understanding in his actions, despite this not being the case. We can call the first of these two kinds of failure to act as an agent deluded determinism, as it is based on an agent's underestimating his capacity to express his self-understanding in his actions, by using his empirical nature as an excuse. In deluded determinism, the agent mistakenly sees himself as wholly determined by forces outside his own control, and as such he ignores the effect his normative judgments have on his behaviour and evades any responsibility he has for his mental states and actions. By refusing to acknowledge the effect of his deliberative conclusions in his actions, the agent acts as if any reasons he might have for acting in the ways he does have no relevance to his actual behaviour, which prevents him from expressing a genuine self-understanding in his actions.

The second kind of self-deceptive inauthenticity we can simply call, taking a cue from Moran, wishful thinking. In wishful thinking, the agent acts as if his normative judgments can have an immediate effect in his actions, regardless of his psychological constitution and his empirical history. The agent does not express a genuine self-concept in his actions in these cases, because he fails to manifest the dispositions that would fit his deliberative conclusions and express his reasons for acting, despite his claims to the opposite. While in cases of deluded determinism the agent underestimates the effect his deliberative conclusions have on his actions, in wishful thinking the agent overestimates this effect and acts as if his actions are the direct result of his normative judgments. These two kinds of self-deceptive inauthenticity correspond to two sides of a conflict between the empirical and the deliberative stance, with deluded determinism signifying an agent's retreat to the empirical standpoint in order to escape the force of his normative judgments and evade taking responsibility for exercising his agency over his actions, and wishful thinking signifying an agent's retreat to the deliberative stance in order to avoid facing his empirical limitations.

In addition to these kinds of agentive failings, I think it's worth drawing attention to another kind of self-deceptive inauthenticity, which we might call rampant rationalization. Rampant rationalization is similar to wishful thinking in that it also involves the use of the deliberative stance to mask one's empirical constraints, but it also seems to involve a more subtle misunderstanding of one's nature than the one present in cases of wishful thinking. Victoria McGeer discusses such a case (which she also identifies as involving rationalization on the agent's part) in the "Moral Development of First-Person Authority", by using the behaviour of Nicholas Bulstrode, a character from George Eliot's "Middlemarch", as an example.[131] McGeer presents this case as a counter-example to Moran's model of ideal agency, in which the agent's deliberative conclusions are the agent's intentional states. As we've seen in chapter 2, McGeer disagrees with this claim and argues that an agent might need to cultivate the dispositions constituting his intentional states through various regulatory means, even after having made the normative judgments that express these states. The case of Bulstrode is meant to illustrate that an agent might seem to function perfectly well in accordance to Moran's model of ideal agency, but still fail to act as an agent because of a problematic self-understanding on his part.

[131] See McGeer, 2007a. For her presentation of the case in the text, see especially pp. 96-98. This example originates in George Eliot, 1996, *Middlemarch*, Oxford: Oxford University Press.

Bulstrode's example also fits our current discussion, as a case of rampant rationalization in which the nature of the agent's flaw in his self-understanding is not as immediately clear as in cases of wishful thinking.

Bulstrode, as presented by McGeer, is an evangelical banker who sees himself as God's faithful servant and whose various motives and actions have one thing in common for him: they are instrumental to furthering God's causes. Acting under this self-concept, Bulstrode's dispositions are all taken by him to fit the judgments constituting his self-understanding, and his deliberative conclusions "are as spontaneous and psychologically effective as anyone might wish who aspires to a condition of rational autonomy."(McGeer, 2007a, p.96). Having developed this peculiar kind of self-understanding, Bulstrode is in a position to offer reasons for all his actions in order to make them intelligible to himself and others, since he can claim that they are all part of his plan to further God's causes and that they all fit his motives which are chosen by God for this purpose. So there is a sense in which our evangelist seems to properly exercise his agency over his actions, as he is able to fit his actions to his self-concept and make these actions intelligible by appealing to their motivating reasons.

What's gone wrong here? Here's the main problem that McGeer identifies with Bulstrode's behaviour:

> "[H]is reason is geared to authorize in him-and for him alone-whatever ambitions and temptations he experiences since these must be connected with God's design. Hence, the appearance of hypocrisy: For he can readily condemn in others the self-same attitudes and actions that he authorizes in himself. In them they are evil and contemptible, whereas in him there is a divine purpose that they ultimately serve." (McGeer, 2007a, p. 98)

What's wrong with Bulstrode is that he is completely unrealistic about himself and he fails to take into consideration the ways in which his psychological constitution is similar to the objects of his contempt. If Bulstrode can be taken as an exemplar case of rampant rationalization, we can argue that in such cases the agent has an empirically unrealistic understanding of himself and his circumstances that allows him to put on a façade of ideal rationality, as it enables him to rationalize all of his actions and attitudes. Rampant rationalization is interesting in that it involves a more skilful kind of self-deception that the mere denial of the empirical facts that wishful thinking seems to involve. The agent in this case doesn't just ignore his psychological constitution, but also consistently confabulates in order to make his actions intelligible to himself and others and create the appearance of well-functioning agency.

This final kind of self-deceptive inauthenticity is also particularly relevant to our present account as it illustrates the importance, for this account, of our common training in the norms inherent in our folk-psychological understanding of rationality, which allow us to determine whether an agent is unrealistic with respect to his self-understanding and the reasons he offers in support of his actions. We can claim that an agent like Bulstrode is completely unrealistic about his circumstances and his nature because we have this common folk-psychological understanding of what it means to act as an agent, and we are in a position to collectively determine whether an agent's reasons make his actions intelligible. I think that McGeer's refinement of Moran's account of agency is compatible with this view, as McGeer herself frequently stresses the importance of folk psychology for our development as agents (see, for example, Chapter 1).

According to our account, acting as an agent involves more than just being able to offer reasons for one's actions, but it also depends on an understanding of what kinds of reasons are appropriate within the social framework in which agents interact. Our training in folk psychology enables us to develop such an understanding, which is informed by an appreciation of the fact that our actions should not only be intelligible to us, but also to the agents we collaborate with. In the case of Bulstrode, what makes his understanding of his actions unrealistic is that he fails to make his actions conform to a common understanding of rationality, because he is inconsistent in his treatment of the same motives he identifies in himself and in others. In the penultimate section of this chapter, I will consider self-deceptive inauthenticity from the perspective of our practices of collective scrutiny of one another's reasons based on our shared standards of intelligibility and examine to what extent we can modify each other's self-understanding while still respecting each other's agency.

Bad reasons and collective criticism

The importance of our shared standards of intelligibility in an account of human agency has been the persistent theme of this thesis. In accordance to our account, the extent to which an agent coherently expresses himself in his actions in an authoritative, self-controlled manner also depends on the agent's capacity to provide intelligible reasons for his actions. The extent to which the reasons an agent has for acting are intelligible depends on the standards of intelligibility this agent shares with others. In human agency, our common folk-psychological understanding of intentionality sets the standards in accordance to which we determine whether an agent's reasons are intelligible. We are in a position to criticize agents who fail to coherently express themselves in their actions, when these agents fail to provide intelligible reasons for their actions. Agents who fail to provide intelligible reasons for their actions might do so because they are unable to provide any kind of justification for their actions, or because the reasons they do appeal to are unintelligible. In the latter cases, we might say that the agents provide bad reasons for acting because the reasons they appeal to don't fit their actions in a way that is intelligible in accordance to our common folk-psychological understanding of agency.

In the case of the man playing the role of the waiter, which fits what we've called deluded determinism, the man fails to act as an agent because he fails to take responsibility for his actions as the products of his own agency. We are in a position to criticize this agent as failing to display a coherent self-understanding in his actions because he fails to provide any reasons for his actions and to take responsibility for them, treating them instead as determined by something other than his own reasons for acting. In doing so, the man also fails to live up to our shared understanding of what is expected of him when he acts as a waiter. We expect his duties and rights as a waiter (recall Sartre's description of this example) to play a role in his decisions to act as one, but our expectation is frustrated by the man's evasion of responsibility for performing these duties and accepting these rights. In this case, the problem is not that the reasons the man has for acting in the way he does are unintelligible to us, but that he does not have any reasons for his actions that he can appeal to in order to make these actions intelligible.

In the case of Bulstrode, which fits what we've called rampant rationalization, he does have reasons for acting in the ways he does. The problem is that these reasons are unintelligible when placed within the context of a folk-psychological understanding of agency. Bulstrode does not manage to make his actions intelligible because he is inconsistent in his justification of these actions. The same motives he justifies in his case as serving God's will, he condemns in the case of others. In doing so, he creates a special kind of explanation for his case which does not respect the ways in which he is psychologically similar to other agents. Bulstrode also frustrates our expectations of intelligible expression of self-understanding on his part, because he fails to understand that the explanations he provides of his motives should also be provided for others with such motives, otherwise it is unclear what the nature of these motives is. If one of his motives is a desire for power, for example, and if he justifies this intentional state by referring to God's will, he should also similarly interpret other agents who are motivated by this kind of state. If he instead finds this desire commendable in himself but despicable in others, then he doesn't seem to share a common understanding of this motive with other agents and as such his reasons for acting which refer to this motive would fail to fit the standards of intelligibility that these agents expect him to share with them. This wouldn't simply be a matter of disagreeing with other agents about the nature of his desire for power. Instead, his understanding of this intentional state seems unintelligible because it cannot be generalized in all cases wherein such a motive is identified. Bulstrode's appeal to God's will as a special explanation for why his own case is special, moreover, doesn't work as it could obviously be used in order to rationalize any kind of motive by him and there is no clear reason why this kind of explanation would not hold for other cases.

Collective criticism of an agent's behaviour, based on this line of reasoning, involves recognizing when an agent fails to make his actions intelligible by failing to provide good reasons for them, because he either fails to provide any reasons for acting in the ways he does or provides reasons that are unintelligible in the context of our shared standards of intelligibility. Agents that persistently act in this self-deceptive manner can be said to display a flawed self-understanding. It seems that these agents' self-understanding can become more genuine, in principle, as long as they arrive at a position to offer more intelligible reasons for their actions, and others can intervene to bring such reasons to their attention. The man playing the role of waiter can express a more coherent self-understanding in his actions as long as he realizes that it’s up to him to carry out the actions associated with his role and that this role does not determine his actions independently of his active contribution in them. Having achieved this self-understanding, the man will more coherently act as a waiter because he will be in a position to offer intelligible reasons for acting in the ways he does and to take responsibility for these actions as the products of his own agency. He can appeal, for example, to the various tasks he has and to his intention of carrying them out, instead of acting as if his behaviour is the automatic product of a process over which he has no control. The self-deceived evangelist might also have the possibility of achieving a more genuine understanding of himself, by recognizing both that it is his desire for power that mainly motivates his actions and that his actions are open to the same kinds of interpretations and criticisms as the actions of agents who are mainly motivated in the same way.

The main conclusion I wish to draw from the aforementioned considerations is that understanding self-deceptive inauthenticity, like understanding self-knowledge and human agency, also depends on understanding the common folk-psychological practices engaged in by us as competent agents. The process of intentional interpretation and evaluation of each other’s actions in the context of our shared standards of intelligibility might enable us to recognize some of the ways in which we fall short of coherently expressing ourselves in our actions. We might then have the capacity to recognize self-deceptive inauthenticity in ourselves and others and to do something about it. In many cases, self-deception might be so entrenched in our behaviour and thinking that nothing short of a complete shift in our circumstances and behavioural patterns can bring us out of acting in an inauthentic manner. Such a shift might frequently be practically impossible to arrange, especially considering the benefits that self-deception might have for us, benefits that we might be implicitly motivated to maintain. Be that as it may, I think there might still be cases in which a more genuine self-understanding might be practically possible for us when we fall victim to self-deception. By changing some of our circumstances and our behavioural patterns, for example, we might arrive at a better position to express a genuine self-understanding in our actions.

We might be able to recognize what changes we need to make ourselves, and take steps to bring them out. In cases of self-deceptive inauthenticity though, it seems more plausible to say that self-deceived agents can only alter their circumstances and behavioural patterns in order to evade their self-deceptive inauthenticity with the help of other agents, who can more easily recognize those agents' self-deception and take steps to counter it. In the case of self-deceived agents who are persistently unable to express a genuine self-understanding in their actions, frequently falling victim to self-deceptive inauthenticity, the best approach might be to take control out of these agents' hands and alter their circumstances and behavioural patterns in the most beneficial way for them. However, a different approach might be to get these agents to endorse the reasons for acting that would lead to a more coherent self-understanding being expressed by them in their actions, and to take the steps needed to express this self-understanding themselves. I don't have a clear answer for which of these approaches works best in general, as I think this would depend on the individual cases encountered and on the different circumstances associated with each. However, I think the benefits derived from adopting the latter approach are worth considering. Attempting to use the first approach in altering the agent's behavioural patterns and circumstances can disrupt his agency as a consequence and leave him unable to express a coherent point of view in his actions in the long-term, which is an outcome that McGeer also warns against:

> "[T]here has always been in some circles, and is perhaps now on the increase, a 'psychiatric model' of human behavior that replaces the structuring ideal of the responsible agent with the notion of a treatable patient-one whose affective responses are debilitating and best controlled by therapy or (increasingly) medication. Of course, in 'fixing' the patient, little heed may be paid to the coherence of her responses to environmental conditions. If the person is indeed responding in a coherent way, it may be wondered how adjusting such responses effects her long-term ability to understand her own experiences as manifestations of a stable and coherent persona. That is, if the person is increasingly directed to attend to her current feelings with an eye to alleviating them, how is she to use her own experiences to build and modify her understanding of the nature (and rationality) of human response to a complex world? Relieved of the need to understand whether her responses are generalizable because they make sense under particular circumstances, she is relieved also of the motivation for challenging those circumstances that give rise to her current experiences." (McGeer, 1996, p. 513, footnote 36)

Conclusion

The question I have started with is this: Can we provide an account of human agency that strikes a balance between our sense of being in control of our actions and our nature as complex, physically constrained organisms? Even trying to articulate what human agency seems to involve immediately led to puzzling concerns. We are now in a much better position to make sense both of our status as agents and of the way this status can be reconciled with the empirical facts of our nature. The crucial idea that structured our discussion is that human agency should be examined as a phenomenon that emerges from our interpersonal interactions and that is essentially linked to these interactions. We cannot make sense of what it means to act as a human agent if we don't also consider the framework in which a human agent acts. We cannot somehow isolate our agency from our social nature and provide a reductive explanation for it that does not take into consideration our collective folk-psychological practices.

Sure enough, explanations that focus on the facts of our empirical nature are essential for providing a well-rounded account of agency, because as we've seen acting as a human agent also involves a specific kind of regulation and we can better understand the limits of our self-control if we understand the limits of our physical constitution. As we've seen, we are self-regulated, language-using creatures that use a self-concept to guide their actions, and our behaviour is largely shaped by complex, distributed processes over which we have limited conscious control. These facts all play a role in understanding how it is that we can succeed and fail to express our agency in our actions. But they are only significant for a full account of agency if placed within the context of our common standards of intelligibility. We can fail or succeed in expressing ourselves in our actions as agents because we train ourselves to act in a manner intelligible not only for ourselves, but for everyone who is similarly trained to understand human thought and action. Our successes and failures in this respect can only be assessed from the perspective of our shared understanding of what acting as a self-knowing agent involves.

The stance we have adopted in this thesis is informed by the idea that our social nature is the key to understanding the nature of our agency. As I hope to have shown, from within this stance, talk of self-knowledge, self-regulation, expressing one's unique point of view in one's actions, reason-guided behaviour and self-deceptive inauthenticity can be significantly disambiguated. As such, this stance should also inform our future investigations on these subjects. Only by understanding more clearly the manner in which we are constrained not only by our physical constitution, but also by our interpersonal interactions and the norms these involve, can we make more sense of our nature and its limits.

Acknowledgements

First and foremost, I wish to thank my primary supervisor, Dr. Tillmann Vierkant, whose helpful suggestions, patient guidance and unwavering stare kept me from self-dissolution and encouraged the right kinds of self-deception that made the completion of this thesis possible. I also owe a debt of gratitude to the members of the audience in my presentations in the University of Twente, Enschede, the Hanse-Wissenschaftscolleg, Delmenhorst and the University of Glasgow, and especially to the staff and colleagues in the University of Edinburgh that have commented on different stages of my project. Special thanks go to Prof. Andy Clark, Prof. Duncan Pritchard, Dave Ward, Jonas Christensen, Diego Zucca, Andy McKinley and Ulla Schmid. Finally, I wish to thank Dr. Victoria McGeer and Dr. Matthew Chrisman for their detailed recommendations that have enabled me to improve the final draft of my thesis.

Bibliography

Bargh, J. A. (2006), "What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior", *European Journal of Social Psychology*, 36, pp. 147-168. [Agenda 2006 article]

Bargh, J. A. (2005), "Bypassing the Will: Towards Demystifying the Nonconscious Control of Social Behavior", in R. Hassin, J. Uleman and J. Bargh (eds.), *The New Unconscious,* New York: Oxford, pp. 37–58.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K. and Trotschel, R. (2001), "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals", *Journal of Personality and Social Psychology* 81, pp. 1014-1027.

Bargh, J. A. and Chartrand, T. L. (1999a), "The Unbearable Automaticity of Being", *American Psychologist* 54, pp. 462-479.

Bargh, J. A. and Chartrand, T.L. (1999b), "The Chameleon Effect: The Perception-Behavior Link and Social Interaction", *Journal of Personality and Social Psychology* 76, pp. 893-910.

Baumeister, R.F., Bratslavsky, E., Muraven, M. and Tice D.M. (1998), "Ego Depletion: Is the Active Self a Limited Resource?", *Journal of Personality and Social Psychology* 74, pp. 1252-1267.

Bruner, J. (1990), *Acts of Meaning,* Harvard University Press.

Bruner, J. (1983), *Child's talk: learning to use language*, Norton, New York.

Bilgrami, A. (1998), "Self-knowledge and Resentment", in C.Wright, B.C.Smith and C.Macdonald (eds.) *Knowing Our Own Minds,* Oxford: Oxford University Press.

Carruthers, P. (1996), "Simulation and Self-Knowledge: A Defense of Theory-Theory", in P. Carruthers and P. K. Smith (eds.) *Theories of Theories of Mind,* Cambridge University Press, pp. 22-39.

Chen, M. and Bargh, J.A. (1997), "Nonconscious Behavioral Confirmation Processes: The Self-Fulfilling Consequences of Automatic Stereotype Activation", *Journal of Experimental Social Psychology*, 33, pp.541-560.

Churchland, P. (1981), "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78, pp. 67-90.

Clark, A. (2007), "Soft Selves and Ecological Control", in D. Ross, D. Spurrett., H. Kincaid and G.L. Stephens (eds.), *Distributed Cognition and the Will,* MIT, pp. 101-122.

Clark, A. (2002), "That Special Something: Dennett on the Making of Minds and Selves", in A. Brook and D. Ross, (eds.), *Daniel Dennett,* Cambridge University Press.

Davies, M. and Stone, T. (2000), "Simulation Theory", *Routledge Encyclopaedia of Philosophy Online.*

Davies, M. and Stone, T. (1996), "The Mental Simulation Debate: A Progress Report", in P. Carruthers and P.K. Smith (eds.), *Theories of Theories of Mind,* Cambridge: Cambridge University Press, pp.119-137.

Dennett, D.C. (2003), "Freedom Evolves", Penguin Books.

Dennett, D.C. (1996), *Kinds of Minds,* New York: Basic Books.

Dennett, D.C. (1993), *Consciousness Explained,* Penguin Books.

Dennett, D.C. (1992), "The Self as a Center of Narrative Gravity", in F.Kessel, P.Cole and D.Johnson (eds.) *Self and Consciousness: Multiple Perspectives,* Hillsdale, NJ: Erlbaum

Dennett, D.C. (1991/2008), "Real Patterns", in W.G. Lycan and J.J. Prinz (eds.) *Mind and Cognition: An Anthology,* Blackwell Publishing Ltd, pp. 323-336.

Dennett, D.C. (1981/2008), "True Believers: The Intentional Strategy and Why it Works", in W.G. Lycan and J.J. Prinz (eds.) *Mind and Cognition: An Anthology,* Blackwell Publishing Ltd, pp. 351-366.

Devine, P. G. (1989), "Stereotypes and Prejudice: Their Automatic and Controlled Components*", Journal of Personality and Social Psychology* 56, pp. 5–18.

Fernyhough, C. (1996), "The Dialogic Mind: a Dialogic Approach to the Higher Mental Functions", *New Ideas in Psychology* 1, pp. 47–62.

Eliot, G. (1996), *Middlemarch*, Oxford: Oxford University Press.

Festinger, L. and Carlsmith, J.M. (1959), "Cognitive Consequences of Forced Compliance", *Journal of Abnormal and Social Psychology* 58, pp. 203-211.

Fitzsimons, G.M and Bargh, J. A. (2003), "Thinking of You: Nonconscious Pursuit of Interpersonal Goals Associated With Relationship Partners", *Journal of Personality and Social Psychology,* 84, pp. 148-164.

Frankfurt, H.G. (1971), "Freedom of the Will and the Concept of a Person", *The Journal of Philosophy 68*, pp. 5-20.

Gallagher, S. (2001), "The Practice of Mind: Theory, Simulation, or Primary Interaction?", *Journal of Consciousness Studies* 8, pp. 83-108.

Gallese, G., Fadiga, L., Fogassi, L. and Rizzolatti, G. (1996), "Action Recognition in the Premotor Cortex", *Brain* 119(2), pp. 593-609.

Goldman, A. (1993), "The Psychology of Folk Psychology", *Behavioral and Brain Sciences* 16, pp. 15-28.

Goldman, A, (1992), "In Defense of the Simulation Theory", *Mind and Language* 7(1), pp. 104-119.

Goldman, A. (1989), "Interpretation Psychologized", *Mind and Language* 4, pp.161-185.

Gopnik, A. (1993), "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality", *Behavioral and Brain Sciences* 16, pp.1-14.

Gopnik, A. and Wellman, H. (1992), "Why the Child's Theory of Mind Really Is a Theory", *Mind and Language* 7, pp. 145-171.

Gordon, R. (1995), "Simulation without Introspection or Inference from Me to You", in T. Stone and M. Davies (eds.) *Mental Simulation: Evaluations and Applications*, Oxford: Blackwell Publishers, pp. 53-67.

Gordon, R. (1992), "The Simulation Theory: Objections and Misconceptions, *Mind and Language* 17, pp.11-34.

Gordon, R. (1986), "Folk Psychology as Simulation", *Mind and Language* 1, pp. 158-171.

Greenwald, A.G. and Banaji, M.R. (1995), "Implicit Social Cognition: Attitudes, Self-Esteem and Stereotypes", *Psychological Review* 102(1), pp. 4-27.

Grusec, J.E. and Redler, E. (1980), "Attribution, Reinforcement and Altruism: A Developmental Analysis", *Developmental Psychology* 16, pp. 525-534.

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

Heal, J. (1998), "Co-cognition and Offline Simulation: Two Ways of Understanding the Simulation Approach", *Mind and Language* 13(4), pp. 477-498.

Holton, R. (2001), "What Is the Role of the Self in Self-Deception?", *Proceedings of the Aristotelian Society*, New Series 101, pp.53-69.

Humphrey, N. and Dennett, D.C. (1989), "Speaking for Ourselves: An Assessment of Multiple Personality Disorder", *Raritan* 9(1), pp.68-98.

Hurley, S. (2006), "Bypassing Conscious Control: Unconscious Imitation, Media Violence, and Freedom of Speech". In S. Pockett, W. P. Banks and S. Gallagher (eds.), *Does Consciousness Cause Behavior?*, MIT Press.

Hutto, D. (2007), "Folk-Psychology without Theory or Simulation", in

D. Hutto, M. Ratcliffe (eds.) *Folk Psychology Re-Assessed*, pp. 115-135.

Hutto, D. (2004), "The Limits of Spectatorial Folk Psychology", *Mind and Language* 19(5), pp. 548-573.

Ismael, J. (Forthcoming), "Selves and Self-Organization", *Minds and Machines.* Available at http://homepage.mac.com/centre.for.time/ismael/

Ismael, J. (2006), "Saving the Baby: Dennett on Autobiography, Agency and the Self", *Philosophical Psychology,* 19(3): 345-360.

Libet, B. (1985), "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action", *Behavioral and Brain Sciences* 8, pp. 529-566.

McGeer, V. (2007a), "The Moral Development of First-Person Authority", *European Journal of Philosophy* 16(1), pp. 81-108.

McGeer, V. (2007b), "The Regulative Dimension of Folk Psychology", in D.Hutto and M.Ratcliffe (eds.), *Folk Psychology Re-Assessed*, Dordrecht: Springer, pp.138-156.

McGeer, V. (2001), "Psycho-Practice, Psycho-Theory and the Contrastive Case of Autism: How Theories of Mind Become Second-Nature", *Journal of Consciousness Studies* 8 (5-7), pp. 109-132.

McGeer, V. (1996), "Is Self-Knowledge an Empirical Problem? Renegotiating the Space of Philosophical Explanation", *Journal of Philosophy* 93*,* pp. 483-515.

McGeer, V and Pettit, P. (2002), "The Self-Regulating Mind", *Language and Communication* 22(3), pp. 281-299.

Mele, A.R. (2001), *Self-Deception Unmasked,* Princeton University Press.

Mele, A.R. (1997), "Real Self-Deception", *Behavioral and Brain Sciences* 20, pp. 91-102.

Metzinger, T. (2003), *Being No One: The Self Model Theory of Subjectivity,* MIT.

Miller, R.L, Brickman, P. and Bollen, D. (1975), “Attribution versus Persuasion as a Means for Modifying Behavior”, *Journal of Personality and Social Psychology* 31, pp. 430-441.

Moran, R. (2001), *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, NJ: Princeton University Press.

Moran, R. (1999-2000), “The Authority of Self-Consciousness”, *Philosophical Topics,* pp. 179-200.

Moran, R (1997), “Self-Knowledge, Discovery, Resolution, and Undoing”, *European Journal of Philosophy* 5 (2), pp. 141-161.

Moskowitz, G. B., Gollwitzer, P.M., Wasel, W., and Schaal, B. (1999), “Preconscious Control of Stereotype Activation through Chronic Egalitarian Goals”, *Journal of Personality and Social Psychology* 77, pp. 167–184.

Muraven, M., Tice D.M. and Baumeister, R.F. (1998), “Self-Control as Limited Resource: Regulatory Depletion Patterns”, *Journal of Personality and Social Psychology* 74, pp. 774-789.

Nichols, S. (1998), “Folk Psychology”, in *Encyclopaedia of Cognitive Science*, London: Nature Publishing Group, pp. 134-140.

Nisbett, R.E. and Schachter, S. (1966), “Cognitive Manipulation of Pain”, *Journal of Experimental Social Psychology* 2, pp. 227-236.

Nisbett, R.E. and Wilson, T.D. (1977), “Telling More Than We Can Know: Verbal Reports on Mental Processes”, *Psychological Review* 84 (3), pp. 231-259.

O’ Connor, T. (1995), “Agent Causation”, in T. O’Connor (ed.) *Agents, Causes and Events: Essays on Indeterminism and Free Will*, New York, Oxford University Press, pp. 173-200.

Pettit, P. (2007),”Neuroscience and Agent-Control”, in D. Ross, D. Spurrett, H. Kincaid and L.G. Stephens, (eds.), *Distributed Cognition and the Will,* MIT, pp. 77-91.

Priest, S. (ed.), (2001), *Jean-Paul Sartre: Basic Writings*, Routledge.

Prinz, W. (2005), “An Ideomotor Approach to Imitation”, in S. Hurley and N. Chater (eds.) *Perspectives on Imitation: From Neuroscience to Social Science* (Vol. 1), Cambridge, MA: MIT Press, pp. 141-157.

Prinz, W. (1990), “A Common Coding Approach to Perception and Action”, in O. Neumann and W. Prinz (eds.) *Relations between Perception and Action,* Berlin: Springer, pp. 167-201.

Rizzolatti, G. and Craighero, L. (2004), “The Mirror-Neuron System”, *Annual Review of Neuroscience* 27, pp. 169-192.

Schachter, S. and Singer, J.E. (1962), “Cognitive, Social and Physiological Determinants of Emotional State”, *Psychological Review* 69, pp. 379-399.

Schwitzgebel, E. (2008), “The Unreliability of Naïve Introspection”, *Philosophical Review* 117, pp. 245-273.

Schwitzgebel, E. (2005), “Acting Contrary to our Professed Beliefs”, available at http://www.faculty.ucr.edu/~eschwitz/

Schwitzgebel, E. (2002), “A Phenomenal, Dispositional Account of Belief”, *Nous* 36, pp. 249-275.

Schwitzgebel, E. (2001), “In-Between Believing, *Philosophical Quarterly* 51, pp. 76-82.

Shah, J. Y. and Kruglanski, A.W. (2003), “Automatic for the People: How Representations of Significant Others Implicitly Affect Goal Pursuit”, *Journal of Personality and Social Psychology,* 84, pp. 661-681.

Steele, C.M. and Liu, T.J. (1983), “Dissonance Processes as Self-Affirmation”, *Journal of Personality and Social Psychology* 45, pp. 5-19.

Stich, S.P. and Nichols, S. (1995), “Second Thoughts on Simulation”, in T. Stone and M. Davies (eds.) *Mental Simulation: Evaluations and Applications,* Oxford: Blackwell Publishers, pp. 87-108.

Stich, S.P. and Nichols, S. (1992), “Folk Psychology: Simulation or Tacit Theory?”, *Mind and Language* 7(1), pp. 35-71.

Swann Jr., W.B., De La Ronde, C. and Hixon, G. (1992), “Embracing the Bitter “Truth”: Negative Self-Concepts and Marital Commitment”, *Psychological Science* 3, pp. 383-386.

Swann Jr., W.B. and Hill, C.A. (1982), “When Our Identities Are Mistaken: Reaffirming Self-Conceptions through Social Interaction”, *Journal of Personality and Social Psychology* 43, pp. 59-66.

Swann Jr., W.B. and Read, S.J. (1981), "Self-Verification Processes: How We Sustain Our Self-Conceptions", *Journal of Experimental Social Psychology* 17, pp. 351-372.

Tetlock, P. (2002), "Social Functionalist Frameworks for Judgments and Choice: Intuitive Politicians, Theologians, and Prosecutors", *Psychological Review* 109, pp. 451–471.

Velleman, J.D., "The Self as Narrator", available at http://www.uwm.edu/~hinchman/Velleman-Dennett.pdf

Velleman, J.D. (2009), *How We Get Along,* Cambridge University Press.

Velleman, J.D. (2007a), "What Good is a Will", in A. Leist *Action in Context*, Berlin/New York: de Gruyter, pp. 193-215.

Velleman, J.D. (2007b), "The Way of the Wanton", in K. Atkins and C. MacKenzie (eds.) *Practical Identity and Narrative Agency*, London: Routledge, pp. 169-192.

Velleman, J.D. (2000), "From Self-Psychology to Moral Philosophy", *Philosophical Perspectives* 14, pp. 349-377.

Velleman, J.D. (1992), "What Happens When Someone Acts?", *Mind* 101, pp.461-481.

Watson, G. (1975), "Free Agency", *The Journal of Philosophy* 72*,* pp. 205-220.

Wegner, D.M. (2002), *The Illusion of Conscious Will*, Cambridge MA: MIT Press.

Wegner, D.M, Wheatley, T. (1999), "Apparent Mental Causation: Sources of the Experience of Will", *American Psychologist* 54, pp. 480-492.

Wimmer, H. and Perner, J. (1983), "Beliefs about Beliefs: Representation and the Containing Function of Wrong Beliefs in Young Children's Understanding of Deception", *Cognition* 13, pp. 103-128.

Zawidski, T.W. (2008), "The Function of Folk Psychology: Mind Reading or Mind Shaping?", *Philosophical Explorations* 11(3), pp. 193-210.

Zillman, D. (1978), "Attribution and Misattribution of Excitatory Reactions", J.H Harvey, W. Ickes and R.F. Kidd (eds.) *New Directions in Attribution Research,* Vol. 2, pp. 335-368.

Zillman, D., Johnson R.C., and Day, K.D. (1974), "Attribution of Apparent Arousal and Proficiency of Recovery for Sympathetic Activation Affecting Excitation

Transfer to Aggressive Behavior", *Journal of Experimental Social Psychology* 10, pp. 503-515.