# How should our conception of agency accommodate the influence of implicit cognition on action?

Naomi Clare Alderson

UCL

*MPhil Stud Philosophical Studies*

2019

*I, Naomi Clare Alderson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.*

# Abstract

Recent study in the field of empirical psychology has revealed the subtle ways in which our actions are influenced by implicit cognition – cognition that typically operates beyond awareness and beyond direct deliberative control. For instance, implicit racial biases make some agents more likely to shoot unarmed black men than unarmed white men in simulations when they are forced to decide quickly (Payne, 2006). Behavioural effects like these have been found even in participants with strong explicit commitments to equality (Monteith, et al., 2001) and by participants making a concerted effort to respond in an unbiased way (Payne, 2006), suggesting that implicit cognition can lead agents to act in ways that they would not choose to. This plausibly suggests that the influence of implicit cognition ought not to be considered agential. However, other studies reveal ways that implicit cognition can help an agent to act as she would choose to, such as the way that it can enable goal-directed habitual action (Snow, 2006) and inhibit the influence of unwanted bias in some agents with egalitarian goals (Moskowitz & Li, 2011). Where implicit cognition seems to help agents achieve their goals and act upon their values, it seems more plausible to think that it is an agential influence on behaviour. This dissertation asks how our theory of agency should accommodate the influence of implicit cognition on behaviour in light of these findings, aiming to determine when – if at all – its influence should be considered part of the kind of agency distinctive of persons.

## Impact Statement

This thesis aims to provide the groundwork for philosophical inquiry concerning how to accommodate the influence of implicit cognition in a theory of action. It could benefit the field of action theory in the following ways. First, it draws attention to the need to better understand the influence of implicit cognition on action in order to have a complete theory of the agency distinctive of persons. If past thinkers have overlooked some aspects of personal agency because they did not consider how implicit cognition influences it, then focusing on the influence of implicit cognition might help to fill gaps in previous accounts. Second, the thesis provides useful criticisms of the works of others in the field, such as John Doris (2015) and Joshua Knobe (2006), which may help future theorists to see past their mistakes and build on their strengths. Finally, the thesis provides an example of how interdisciplinary approaches can benefit the field of action theory, since it draws on both psychological and philosophical findings to bring to light unsolved problems. To bring about these benefits within academia, excerpts from this work (or the future research that it leads to) might be published in academic journals and presented at conferences, in order to share the insights contained here more widely.

Research into the topics addressed here might also bring about benefits beyond academia. A wide range of institutions are currently working out how to respond to evidence of the pervasive influence of implicit cognition on behaviour, such as the evidence that implicit biases affect a wide range of professional judgements such as medical diagnoses and hiring decisions. Many employers now provide implicit bias training and there is increasing pressure on legal systems to better account for how implicitly-governed behaviour should be treated by the state. Theoretical work that investigates the extent to which implicitly-governed behaviour is agential could help to influence thinking about implicitly-governed behaviour in the workplace, in public policy and in the law. These benefits could be brought about by sharing the insights of research into this question in relevant journals beyond the discipline of philosophy or in mainstream media publications. It could also provide the impetus for future interdisciplinary work, drawing together psychologists, legal theorists and policymakers in order to understand what the societal implications of implicitly-governed behaviour really are (and what they really should be).

# Table of Contents

## Introduction

In recent years, psychological studies in a range of different domains have provided evidence supporting what are known as 'dual-process' theories of reasoning, decision-making and social cognition.[1] Though there are different kinds of dual-process theory, all share a commitment to the idea that human judgement and behaviour are influenced by two different kinds of process: type 1 processes that are fast, automatic and unconscious, and type 2 processes that are slow, controlled, and conscious (Evans, 2008). For example, one might have an intuitive sense that a new acquaintance is not trustworthy without ever having consciously reflected upon it and not knowing why one feels that way. This would be a judgement arrived at using type 1 processes. Alternatively, one might carefully consider whether a new acquaintance is trustworthy, consciously recalling her past behaviour and conducting effortful, sequential reasoning to decide that she is not. This would be a judgement arrived at using type 2 processes. As a shorthand, type 1 and type 2 processes are here referred to using the terms of art of 'implicit cognition' and 'explicit cognition' respectively, where these terms are taken to apply to a broad range of cognitions that exhibit the features listed above to a greater or lesser degree.

Beyond sharing a commitment to two kinds of process, dual-process theories differ substantially. For instance, while some theorists argue that the two kinds of process suggest the existence of two separate cognitive systems, others focus only on the differing processes without presuming that they arise from multiple systems.[2] Further, there is disagreement about how the processes interact. Some theorists suggest that processes run 'in parallel' and sometimes compete, while others maintain that they operate sequentially with implicit cognition preparing a rapid, default response that can be intervened in by explicit cognition (Evans & Stanovich, 2013, p. 227). In addition, the evidence suggests that there are a range of different types of implicit cognitive process that develop and operate in different ways (Evans, 2008). For instance, some implicit cognition is best understood as an automatized process developed from previous frequent explicit cognition – a kind of habituation that Evans describes as the 'automation of thought' (2008, p. 261). On the other hand, some kinds of associative learning involve entirely implicit learning, so that what is

[1] For a review, see Jonathan Evans (2008).

[2] For an example of the two-system view, see Keith Stanovich, 2004. For a discussion of the processes as merely processes, see Evans, 2013.

learnt influences behaviour despite never being the subject of conscious, explicit cognition (Evans, 2008, p. 271). This suggests that a neat categorisation of type 1 and type 2 processes may be misleading if it fails to capture the diversity of implicit cognition.

Traditional conceptions of agency emphasise the influence of explicit cognition on action, focusing on the way that the agency distinctive of persons involves, directly or indirectly, conscious awareness and practical reasoning. For instance, the conception of agency arising from the work of Donald Davidson (2001b) and Elizabeth Anscombe (1963) characterises full-blown agency as intentional action where an intentional act is something that is, under at least one description, known and performed due to a belief that the act will help to satisfy some desire. This model accepts that intentional action need not proceed from conscious deliberation. However, it does require that an intentional action is consciously known to the agent, at least under some description. Further, it requires that intentional acts are performed according to reasoning that the agent would accept if she were aware of it, suggesting that it must at least align with some explicit reasoning, even if it is not directly controlled by it. Alternatively, philosophers such as Christine Korsgaard (2008) argue that full-blown agency is acting for what one takes to be a reason, suggesting that the reasons for acts must at least be possible for agents to call to mind and recognise as reasons. Finally, the work of Harry Frankfurt characterises full-blown agency as those acts which are performed according to the desires that the agent reflectively endorses, suggesting at least indirect governance by explicit reflection (1971).

On the other hand, while much philosophical work has been done to provide models of implicit cognition,[3] to explore its role in virtue,[4] and to establish what (if any) moral responsibility we bear for its influence,[5] relatively little work has been done to explore the ramifications of the influence of implicit cognition for traditional conceptions of full-blown agency. One reason for this is the way that its influence on behaviour seems, at first glance, to be non-agential. For example, one much discussed kind of implicit cognition is implicit bias. Implicit biases are associations between

[3] For an overview, see Jules Holroyd (2016).

[4] See, for instance, Nancy Snow, (2006) and Michael Brownstein and Alex Madva (2012).

[5] See, for instance, Holroyd (2012), Brownstein (2016) and Angela Smith (2018).

certain groups (such as women or black people) and certain characteristics (such as submissiveness or aggression) that can influence action in ways that are typical of implicit cognition: unconsciously, without deliberation, and sometimes in ways that conflict with the agent's explicit attitudes about what to do. For instance, implicit racial biases make agents more likely to shoot unarmed black men than unarmed white men in simulations when forced to decide quickly (Payne, 2006), and they are thought to explain some kinds of discrimination in hiring decisions (Dovidio & Gaertner, 2000). Biased behaviours like these have been found even in participants with strong explicit commitments to equality (Monteith, et al., 2001) and by participants making a concerted effort to respond in an unbiased way (Payne, 2006). Given that this kind of influence of implicit cognition on behaviour can be unknown, does not arise directly from deliberation, and is sometimes counter to agents' reasoning, it first appears to be a non-agential influence on behaviour that need not be accommodated by a theory of agency. It appears more like an unwanted reflex reaction than an instance of full-blown agency.

However, there are two reasons not to accept the conclusion that the influence of implicit cognition should not be accommodated by a theory of the agency distinctive of persons. Firstly, the psychological record suggests that implicit cognition influences a wide range of our actions. If that is the case, a thorough account of agency ought to have something to say about what role implicit cognition plays, since it so frequently contributes to what we do. Secondly, not all instances of the influence of implicit cognition seem so clearly non-agential. Some studies have revealed examples of implicit cognition that do serve the agent's explicit goals, such as implicit cognition that helps agents avoid temptation (Fishbach & Shah, 2006), and cognition that inhibits the influence of unwanted bias in some agents with egalitarian goals (Moskowitz & Li, 2011). Goal-directed behaviour seems importantly different from other behaviours that do not count as instances of the agency distinctive of persons, such as reflex reactions. While the former seems like a candidate for something done by the agent herself, the latter seems like something only her body does. Whether or not the right kind of link to an agent's values and goals is sufficient to make implicitly-governed behaviour an example of full-blown agency deserves careful thought. Preparing the way for a convincing answer to this question is the goal of this dissertation.

In the first chapter, a range of cases of the influence of implicit cognition on action are presented in order to illustrate the potential problems it causes for traditional conceptions of agency. Considering these cases uncovers the intuition central to the rest of the investigation: the intuition that implicitly-governed behaviour seems agential when it is best explained by the longstanding values and goals of the agent, despite lacking some features typically thought to characterise full-blown agency. Following the central intuition, the second chapter evaluates an account of agency provided by John Doris that places alignment with an agent's values at the heart of agency (2015). If Doris's account were successful, it could provide a principle by which to accommodate implicitly-governed goal-directed behaviour within the scope of agency. Unfortunately, Doris's valuational account is too vulnerable to criticism to provide a satisfactory theory of action, suggesting a better model must be developed. By placing values at the centre, Doris's account strays too far from our intuitions about everyday intentional action, leading to possible trouble for the central intuition.

In the final chapter, further trouble is caused by a possible deflationary response to the intuition. The work of Joshua Knobe suggests that intuitions about agency are systematically influenced by normative considerations arguably unfit to provide evidence of what kinds of action there are (2006; 2009). If this analysis proves correct, the intuition that values matter for agency might be too suspect to inform an account of agency. Ultimately, however, the deflationary response is rejected since the findings motivating Knobe's account are found to be compatible with both the central intuition and its evidential value for a theory of agency. Ultimately, it is concluded that the best way to accommodate the influence of implicit cognition on action is to develop a new theory of agency, one that can explain why values seem to matter for agency in some cases but not others. What is needed is a theory of agency that recognises that implicit cognition is sometimes a tool used by human agents to achieve their goals, without implying that other more paradigmatic cases of agency should not be seen as agential.

## Chapter One: Implicit Cognition in Action

Both reflection upon experience and the findings of empirical psychology demonstrate that action is influenced by implicit cognition in a range of different ways. When we act from habits so ingrained that we need not deliberate, initiate action, or fully attend to what we do, it is implicit cognition that guides us. When we are forced to make split-second decisions and cannot consciously deliberate about what to do, it is implicit cognition that (for better or worse) shapes our response. And when we find ourselves faced with the opportunity to achieve a goal, an opportunity we have not consciously recognised as one, it is implicit cognition that drives our goal-pursuing behaviour, even beyond our own awareness.

In this chapter a general model of how some kinds of implicit cognition influence action is sketched, one that draws on a review by psychologists John Bargh and Peter Gollwitzer (2005). While there are a range of different models of implicit cognition, the work of Bargh and Gollwitzer is most closely concerned with how it shapes outward behaviour, making it the most likely model to underpin the effects pertinent to this investigation. Next, three contrasting examples of implicit cognition's influence on action are compared in order to establish where the tension between implicitly-governed action and traditional accounts of action really lies. A brief case is made that only the final case here considered – that of what is here termed autonomous goal-control – is really in tension with traditional conceptions of agency. This tension arises due to contrasting intuitive responses. On the one hand, the fact that the behaviour is best explained by an agent's longstanding goals and her past pursuit of them makes the behaviour appear agential. On the other hand, the way that the implicitly-governed behaviour is unknown to her when she acts and out of reach of her direct deliberative control suggests it is not a proper example of full-blown agency, according to traditional accounts.

## *1.1 The influence of implicit cognition on action: a model*

Though it is here assumed that different implicit processes influence action in different ways, there seems to be a fairly common structure to their influence, as outlined by Gollwitzer and Bargh (2005). First, one or some of the agent's mental representations, such as her judgements or goals, become associated with certain situational features. For instance, repeated exposure to a particular racial stereotype might lead to the association of the judgement that 'that person is aggressive' with the situational cue of a black person's face. Alternatively, repeated efforts to compete with your older sister might lead to the association of competitive behaviours with the presence of (or even thoughts of) your older sister. This association is formed by the repeated pairing of the cue and the mental representation. One way that this repeated pairing might occur is through repeated explicit cognitive processes. For instance, goals chosen through explicit, reflective cognition can, if pursued habitually enough, give rise to implicit processes that serve them (Stanovich, 2004, pp. 66-67). However, it also appears that such associations can be made in one's implicit cognition without conscious awareness and even in conflict with one's explicitly gained beliefs (Evans, 2008).

The more often the pairing between the situational cue and the mental representation occurs, the more 'accessible' the mental representation becomes (Rees & Webber, 2014). The accessibility of a mental representation is a measure of how quickly and how easily it can be ‘activated’, where activated representations are understood as those that are influencing cognition. The more accessible a mental representation is, the more easily it is activated by situational cues (Snow, 2006, pp. 555-556). Every mental representation has a level of chronic accessibility, which is increased by repeated activation. However, encountering situational cues associated with the representation can also increase its accessibility temporarily due to its likely appropriateness to the situation (Rees & Webber, 2014). For instance, repeated attempts to compete with one’s sister might induce an association between mental representations such as ‘I must compete’ and the presence of or thoughts of one’s sister. If this association develops a high chronic accessibility through repeated pairing over time, then the representation ‘I must compete’ and the behaviours it prompts will be quickly and easily activated by situational cues to do with one’s sister. If such a cue is encountered, it will temporarily increase the accessibility of competitive mental representations even further, making it likely that they will influence cognition and behaviour.

Once an association is sufficiently established, exposure to relevant situational cues can non-consciously activate the mental representation, leading it to influence judgements and behaviours in ways that are beyond the agent’s awareness and direct deliberative control (Gollwitzer & Bargh, 2005). For instance, the activation of an implicit racial bias might influence an agent’s judgements about a potential employee in ways that she is unaware of and may struggle to exert direct deliberative control over. This influence is what the influence of implicit cognition on action is here understood to amount to.

While this model is able to account for the influence of implicit cognition most pertinent here, it should not be taken as the only way to explain implicit cognition’s influence. To do so is to overlook the diversity of implicit cognition. Furthermore, there is considerable diversity even within the class of implicit cognitive processes that this model does explain. For instance, the influence of implicit processes considered here varies in the extent to which it lies beyond awareness and direct deliberative control. Its influence on habitual action, for instance, may sometimes be such that conscious effort can override implicit cognition and take back full deliberative control. Cases in which implicit cognition brings about behaviour that is not known under any description, however, does not seem so easy to override. A further way in which the influence of implicit cognition can differ is the way in which it can align with or depart from an agent’s explicit judgements. Implicit cognition’s influence can be in line with the agent’s explicit beliefs and goals, such as when the presence of one's older sister encourages competitive behaviour when one does in fact believe that one should compete with her. Alternatively, its influence can be 'belief discordant' (Brownstein & Mavda, 2012), as when a racial bias that one does not believe in influences one's reaction to a person of that race.

### *1.2 Habitual action*

The most familiar way in which implicit cognition influences action is the way that it guides aspects of habitual action. Habitual actions are understood here, following Snow (2006), as those actions that an agent performs so often over time that they become routinized: they can be performed without deliberation and with only minimal awareness and control. Snow gives the example of driving home by a familiar route in ordinary conditions – what she describes as driving home on 'automatic pilot' (2006, pp. 551-552). When driving on autopilot, a driver can take turns, respond to other traffic, adjust her speed and so on without deliberating about, for instance, whether and when to turn or whether to speed up or slow down. She also drives with only minimal awareness of the acts that constitute her driving and the reasons motivating them. Though aware that she is driving home, she may not be consciously aware of many of the situational cues that she is responding to, or of her responses to them. She may not consciously register the brake lights of the car in front and her own braking in response, she may not notice at all the small adjustments she makes to stay in the middle of her lane, and it is doubtful that she is occurrently aware of the fact that she is taking this particular route because she desires to get home as quickly as possible.

The influence of implicit cognition on action provides an explanation of why habitual actions come to have these distinctive features. Since habits are formed by repeatedly performing the same actions, usually in similar situations, features of the situations in which the habitual act is performed become associated with mental representations and actions linked to the habit. For instance, a Catholic who has the goal to always make the sign of the cross before prayer will come to associate situational features common in the moments before prayer, such as hearing others say 'Let us pray', with the goal to make the sign of the cross and the action of making the sign of the cross. A driver on her usual route home will come to associate the features of a particular turning, such as that house on the left, with the goal of getting home quickly and the actions of braking and turning the wheel. If the mental representations associated with the habit become sufficiently accessible over time then they come to be non-consciously activated by relevant environmental triggers, influencing cognition and action in such a way as to perform the habitual action with only the governance of implicit cognitive processes.

Ordinarily, the implicit cognition guiding habitual action runs autonomously until the habitual action is completed – for instance, guiding the hand in making the sign of the cross without full awareness and conscious control of the agent (Bargh, 2005; Stanovich, 2004). However, its influence may be inhibited by conscious effort or the presence of action-inhibiting situational cues, such as the perception of a child running into the road or the awareness that the person saying 'Let us pray' is a character in a television programme. Further, the agent may intervene in habitual action, snapping back to full awareness and direct deliberative control. This is an example of what Snow describes as 'intervention control' (Snow, 2006, p. 549) – the kind of control we have when we can, through conscious effort, redirect or stop a habitually-performed action. If the agent remembered, for instance, that she needed to visit the shops on her way home then she might exert intervention control over her driving, taking back conscious, effortful control instead of relying on implicit cognition. She might, for example, consciously deliberate about the best route to the shops, pay full attention to the environment in order to make the right turns, and take fuller conscious control of her actions of indicating, changing speed and turning the wheel.

The influence of implicit cognition on habitual action, then, typically has the following features. Firstly, habitual action is not typically fully beyond an agent's awareness, though aspects of it may be. In standard cases of habitual action the agent knows that she is performing the action under some description, such as the description 'driving home', even if she is unaware of many of the situational cues and responses that partly constitute her driving. Secondly, habitual actions are typically performed without deliberation. Instead, a situational cue non-consciously activates the relevant mental representations and behaviours, setting the habitual action into motion and guiding it autonomously to completion. Thirdly, habitual actions are under intervention control – we can take back the reins from our implicit cognition in order to guide action with effortful awareness if we need to or if we so choose.

### *1.3 Situational control*

Situational control is the influence of implicit cognition on action that has motivated a kind of scepticism about agency, where agency is understood as the ability to act according to one's reflectively endorsed beliefs and desires and situational control is presented as a pervasive way in which our actions are not so governed (Doris, 2015).

In a sense, 'situational control' is an ambiguous term to use here: it could be suggested that all influences of implicit cognition on action discussed here involve some kind of situational control, insofar as situational features non-consciously activate cognition that influences behaviour. A more accurate name, perhaps, would be *mere* situational control, as this name reflects the way in which the influence on action in these cases is governed primarily by the situation and not, even indirectly, by the agent and her attitudes. However, in the interests of simplicity and brevity 'situational control' will be used throughout.

In instances of situational control, situational features that have become associated with certain mental representations non-consciously activate those representations. Once activated, the mental representations influence cognition and behaviour in ways that are both *fully* beyond awareness at the time they operate and that cannot be intelligibly explained by reference to the agent's explicit attitudes, such as their beliefs, values and goals. Some cases of situational control are marked by a simple absence of rationalising attitudes: in these cases, it does not seem plausible that the agent has any explicit attitude that would explain the behaviour. In other, perhaps more worrying cases, situational control appears to influence action in a way that the agent would clearly reject, running counter to her explicit beliefs about what to do.

The clearest evidence for the effects of situational control come from studies that involve unconscious 'priming': experiments in which psychologists deliberately non-consciously activate some mental representation and then measure its influence on behaviour (Gollwitzer & Bargh, 2005). For instance, participants may be shown an image for a time so short that they do not consciously perceive it, or they may be presented with words associated with a particular behaviour in a word-sorting activity. By presenting the prime in a way that is unable or unlikely to draw conscious attention, experimenters aim to ensure that any activation of an associated mental representation is non-conscious, rather than being the result of a conscious choice or inference. Further, the experiments are designed in such a way as to ensure that the influence of the activated representation must operate beyond awareness, and/or they use self-report in order to measure whether agents took it to be influencing their behaviour. For instance, in experiments such as those using Implicit Association Tests the activity used to measure the influence of implicit cognition on behaviour must be completed in time-spans that are too short for explicit decision-making, forcing implicit cognition to guide behaviour. In others, though the activity occurs at a slow pace with conscious deliberation, interviews following the activity are used to try to establish whether explicit cognition or implicit cognition was responsible for the behaviour. It is assumed that if the participant did not notice the prime and does not identify the primed mental representation as driving their behaviour then it exerted its influence implicitly.

Considering two experimental examples can help to make the process clearer. In the first study (Bargh, et al., 1996), some non-African American participants were non-consciously primed with photos of black faces while they were concentrating on a shape-sorting task, while others were primed with images of white faces. The images were shown to them for such brief intervals that they were not consciously perceived. After a while, the computers on which participants were working were made to 'crash' while experimenters measured the level of hostility shown in response. A significantly greater level of hostility was shown by those who had been primed with black faces than those who had been primed with white faces. According to the psychologists' interpretation of this effect, the increased hostility was attributable to the implicit association of black faces with a mental representation of hostility, a representation that influenced agents to be more hostile once activated. The pattern was observed even in agents whose self-reported levels of racial bias associating black people with hostility were low, those who would presumably reject both their association between blackness and hostility and the process of seeing a black face as a reason to be more hostile. In the second study (Bargh, et al., 1996), young and healthy participants were given a word-unscrambling task that, for some participants, included lots of words to do with elderly infirmity. Those who had been primed with words to do with infirmity walked significantly more slowly when leaving the experiment. In this case, the unconscious prime influenced behaviour despite the fact that agents presumably lacked any attitude that rationalised it, given the assumption that no agents believed reading words to do with infirmity gave them reason to walk more slowly.

Though it is easiest to detect in experimental conditions, there is evidence that situational control operates in ordinary circumstances too. A non-experimental instance of situational control in which there does not seem to be an attitude that could explain the behaviour is plausibly provided by ballot order effects. Multiple studies of voting habits reveal that being placed first on a ballot significantly increases the number of votes the candidate gets (Marcinkiewicz, 2014; Webber et al., 2014). One way to explain this would be to assume that the situational feature of being first on the ballot non-consciously activates some mental representation, something like the judgement 'The first one is the best one'. Once activated, the mental representation influences the act of choosing a candidate in ways beyond the agent's awareness, making her more likely to choose the candidate listed first. In cases where this occurs, it seems likely that the agent simply lacks any explicit attitude that would rationalise the behaviour, such as the belief that 'The candidate listed first is the best one', or any attitudes that would entail such a judgement in this situation. Though it is certainly possible that some voters *do* choose based on a belief that the candidate listed first is the best choice, it is perhaps more plausible that most voters lack such an attitude and that this is a case of mere situational control.

Cases of what has been termed 'aversive racism' provide examples of situational control in which an agent not only lacks attitudes that would rationalise her implicitly-governed behaviour, but has explicit attitudes that would prohibit such behaviour, so that the behaviour in some ways runs counter to her explicit beliefs (Pearson, et al., 2009). For instance, an agent with the egalitarian explicitly-gained belief that all races are equal may yet have implicit racial biases. Her implicit bias may consist in an association between racial features and some mental representation, such as an association between a black face and a mental representation with content like 'That person is aggressive'. Where this implicit bias influences behaviour, the egalitarian may be led to act in ways counter to her explicit goals, for instance choosing to sit further away from a black person than a white person (McConnell & Leibold, 2001) or rating the application of a candidate who belongs to the Black Student Union more poorly than the equally good CV of a candidate who belongs to an overwhelmingly white fraternity (Dovidio & Gaertner, 2000).

To recap, then, situational control is the influence of implicit cognition on action when that influence is fully beyond the agent's awareness, so that she does not know how her implicit cognition is shaping her behaviour under any description. Participants in unconscious priming studies, for instance, typically do not report awareness of the prime or the way that it influenced their subsequent behaviour, and often even deny the psychologist's explanation of their behaviour if told (Gollwitzer & Bargh, 2005). Presumably due to this lack of awareness, agents seem unable to exert even intervention control over their situationally-controlled actions: since the influence of the cognition lies fully beyond the agent's awareness she cannot intervene in it while it is taking place and snap back to full attentive control. This may explain why even participants told to suppress their bias often fail to do so (Payne, 2006). Finally, while many instances of habitual action can be explained by reference to the agent's goals and beliefs, situation control occurs when the influence of implicit cognition cannot be intelligibly explained or rationalised by reference to the agent's explicit attitudes.

### *1.4 Autonomous goal control*

The final example discussed here is the one that causes the most trouble for traditional conceptions of agency: that of autonomous goal control. Autonomous goal control is a term used here to refer to cases in which implicit cognition helps agents to attain a longstanding goal, though the goal-pursuing behaviour and the implicit cognition guiding it both remain beyond the agent's awareness. Autonomous goal control is one of a class of overlooked cases that would benefit from further study (as argued by Brownstein and Madva (2012) and Holroyd (2016)): cases of behaviour in which the influence of implicit cognition is in line with the agent's explicit beliefs about how she ought to act.

What is so interesting about these cases is the way that they pull our intuitions in different directions. On the one hand, the influence of implicit cognition and the behaviour it governs are typically beyond conscious awareness at the time they occur. This suggests that, like a heartbeat or a digestive process, it should not be considered an instance of agency. However, the goal-pursuing behaviour is part of the way that the agent strives to attain a goal – often one that really matters to her as a person. The behaviour both arises from and demonstrates the value that the agent attaches to it. Moreover, reference to her goal is needed to give a full explanation of the action and make it intelligible when it would, in many instances, otherwise remain puzzling. These features prompt the intuition that the action might be an instance of full-blown agency after all, since what is done arises out of a goal that matters to the agent and her past striving for it. Unlike a heartbeat, the acts governed by autonomous goal control seem like something that the agent does herself.

Early and influential investigation into autonomous goal control was conducted by Gollwitzer and Bargh. In a series of studies (Gollwitzer & Bargh, 2005), these psychologists show that non-conscious priming of an agent's chronically-held goals elicits behaviours that are in pursuit of the goal, guiding behaviour flexibly and responsively towards attaining the goal's end in the same way that a conscious goal would do. For instance, in one study participants who were given a word-unscrambling task that contained lots of words to do with cooperation were significantly more likely to use a cooperative than a competitive strategy in a game that could lend itself to either a cooperative or competitive style of play. Unlike those participants who were explicitly instructed to cooperate, there was nearly no correlation between the amount of cooperation demonstrated in game play and the extent to which agents reported that they had been cooperative, providing substantial evidence that many agents were unaware of the influence of their implicit cognition.

By itself, this study appears to do little more than show further evidence of situational control. All that appears to differ between a study in which words to do with cooperation make an agent unwittingly more cooperative and one in which words to do with the elderly make an agent unwittingly more slow is that being cooperative in a game at least appears warranted from a third-person perspective. This difference seems of no importance here, however, since the appropriateness of this implicitly-governed behaviour may only be a happy coincidence, and there is no reason to think that the cooperative behaviour is an act that the agent would reflectively endorse if she were aware of it.

However, further studies more plausibly track the way that autonomous goal control can provide a way for agents to ensure that their acts are in line with their explicit attitudes goals and beliefs about what they ought to do. For instance, Fishbach and Shah (2006) have presented the findings of five studies in which implicit cognition contributes to successful goal pursuit in the face of temptation to do otherwise. In these studies, agents demonstrate implicitly-governed goal-pursuing behaviour, such as responding more quickly when pushing a button to avoid a tempting stimuli and pulling a lever to engage with the pursued end than vice versa. The effect of the implicit cognition on behaviour was strongest in those who had stronger explicit goals to avoid the temptation, demonstrating a link between agents' judgements about what they ought to do and the influence of autonomous goal control on their behaviour. As Snow notes, studies like these suggest that non-conscious goal activation 'can counteract situational control and promote the personal control of action in accordance with a person's values and priorities' (2006, p. 548). Instead of being subpersonal responses to situational triggers, these examples seem more like implicitly-governed methods of behaving as one believes one ought.

Some of the best examples of autonomous goal control are provided by studies of the control of unwanted implicit bias. For instance, Moskowitz et al. (1999) demonstrate that agents with chronic egalitarian goals – agents who have pursued the goal of treating others fairly persistently over time – are able to pre-consciously inhibit the effect of unwanted implicit bias on their responses in Implicit Association Tests. In a further study, Moskowitz and Li (2011) demonstrate that in situations in which egalitarian goals are activated by, for instance, contemplating failure to treat people of different races equally, they continue to have non-conscious influence on behaviour by inhibiting the influence of racial bias on subsequent Implicit Association Tests. Multiple studies, such as one performed by Webb et al. (2010), have demonstrated the efficacy of using implementation intentions (consciously formed intentions that specify what action will be taken if a specific cue is encountered, such as ‘When I see a woman, I will think leader!’) to establish implicit cognition that effectively inhibits unwanted bias in Implicit Association Tests. Finally, Glaser and Knowles (2008) demonstrate how an implicit motivation to control prejudice pre-consciously inhibits the influence of unwanted implicit bias.

In each study above, implicit cognitive processes inhibit the influence of unwanted implicit bias, helping the agents to achieve egalitarian goals. In cases in which these egalitarian goals plausibly reflect the explicit attitudes of the agent, these are cases in which autonomous goal control helps agents to act in line with their beliefs, goals and values. Further, since there is evidence that conscious, effortful control of unwanted implicit bias can actually be counterproductive,[6] causing rebound effects in which an agent behaves in a more biased way immediately after she stops trying to be unbiased, autonomous control may actually provide a more effective way of acting in line with one’s beliefs than direct deliberative control since it is not so cognitively demanding that it brings about a rebound.

In many cases, it may be difficult to know whether the influence of implicit cognition on action is a case of mere situational control or a case of autonomous goal control. One may wonder, observing the behaviour, whether it really reflects the longstanding goals of the agent or if it is just a brute response to some situational cue – a response not arising from any previous explicit cognition. However, it is here argued that autonomous goal control can be identified when the influence of implicit cognition on behaviour is best explained not only by reference to the situational features that triggered it but also by reference to the explicit attitudes of the agent and her history of acting in line with those attitudes.

Consider the actions of a chronic egalitarian living in a racist society, a society that associates black faces with aggression. Though she has an unwanted implicit bias, she frequently strives to treat people equally in her explicitly-governed behaviour. She often contemplates past failure to treat people fairly and the likelihood that her behaviour is not perfect. Her persistent pursuit of egalitarian goals makes them highly accessible, while her reflection upon past and possible failings directly activates them temporarily after each bout of reflection. Imagine she completes her weekly shop in hurry, distracted by worries about work. In this situation, most people with her

[6] For discussion of various studies, see Moskowitz and Li (2011).

implicit bias would be expected to respond in a biased way to black employees, making less eye contact with them or standing further away from them than their white colleagues. If the chronic egalitarian implicitly appreciates that this is a situation in which she might be biased due to relevant situational cues, and if her goals implicitly influence her to respond in an unbiased way, then her behaviour would provide a clear case of autonomous goal control. Her goal to be egalitarian would, in this instance, influence her to act in an egalitarian way without her conscious guidance or awareness. Furthermore, it would be puzzling without the explanation provided by her explicit goals and past pursuit of them, since most people with her bias would have acted differently.

Though this example is hypothetical, there are good reasons to think it a real and common phenomenon. Since the chronic accessibility of a goal is established by repeated pursuit over time (Gollwitzer & Bargh, 2005), some of the most accessible goals an agent has are the ones that she pursues most often. In some cases this will be because they are the goals she values the most. In many situations, an agent's most chronically accessible goals can be expected to guide her behaviour. If this analysis is correct, it appears that agents have a kind of indirect, long-range control over the influence of their autonomous goal control. What autonomous goal control they have is shaped by the goals they most often strive for – often the things that really matter to them. By consciously striving to act in a particular way, implicit cognition supporting that behaviour is established that later guides behaviour towards the agent's valued end in ways that are beyond the agent's conscious awareness and control.

## *1.5 Accommodating the influence of implicit cognition on action*

The remaining question for this chapter is the question of what traditional conceptions of agency ought to say about each of the examples described above: habitual action, situational control, and autonomous goal control. While there seems to be a straightforward way to include habitual action in a traditional conception of agency, as well as clear reasons to exclude situational control, the best response to autonomous goal control is as yet unclear. Not only does autonomous goal control not fall cleanly within or without a traditional conception of agency, it also raises questions about what is at stake when defining full-blown agency.

Before exploring the three examples and their place within a conception of agency, it is useful to sketch what kind of agency is up for debate. There is, of course, a sense in which all three cases are examples of agency. Agency understood in a broad sense as bodily movement initiated by the body could include unintended and uncontrolled movements such as reflexes, movements made in sleep and so on. This broad conception would clearly include all implicitly-governed behaviour. However, most philosophical accounts of agency investigate not this broad understanding of action but instead a narrower conception: full-blown agency understood as the distinctive kind of action that persons do. For instance, one of the most influential ideas in action theory is the idea that agency finds its full or sole expression in intentional action. Donald Davidson remarks, for example, that 'a person does, as agent, whatever he does intentionally' (Davidson, 2001b). In a nutshell, the motivation behind this idea is this: what we are interested in when we talk about agency is the actions of persons, not bodies or organs, and we are best able to attribute actions to persons when it is something they intend. John Hyman characterises what he calls this 'orthodox doctrine' as follows: what makes 'the contraction of my heart' different from the 'contraction of my fist when I clench it' is the fact that the latter is done intentionally, a difference that matters because intentionality allows an act to be 'imputed to the whole person as an agent' (2015, p. 25).[7] Another account of the distinctive agency of persons is defended by philosophers such as Korsgaard (2008), who argues that what is distinctive of the agency of persons is that their actions are done for what they take to be a reason. It is this narrower kind of agency that is under consideration here – the agency distinctive of persons. What must be asked of each example in turn, then, is not merely 'Is this an action?' but instead 'Is this an action that is of the kind distinctive of persons?'.

The route by which habitual action can be classed as agential is clear. Typical instances of habitual agency are intentional since they are known under some description (such as the description 'I am driving home') and, often, are rationalised by the belief that they will help satisfy some desire, want or urge (such as the desire to get home as quickly as possible). Even if aspects of the action are beyond the agent's awareness as she acts, this is common to most actions. As philosophers following the work of Anscombe (1963) point out, things like the muscle contractions we make as we act are typically unknown and not the object of a specific intention when we act,

[7] This doctrine does not imply, however, that whatever is done as a person is done intentionally. It may well be that some unintentional acts are instances of the agency distinctive of persons – some might describe the implicitly-governed set of goal-serving acts described here as such, for instance. The doctrine is merely invoked here as a way of marking out the full-blown agency distinctive of persons.

but so long as there is some description under which the action is known and intended it can be considered intentional. Moreover, though the belief that her habitual action will help satisfy some desire – or the reason for which she takes herself to be acting – may not be before the agent's awareness when she acts, it is plausible in cases of habitual action that she could uncover it after a little reflection. If asked why one takes a particular route home, for instance, it seems likely one would be able to explain one's desire to get home quickly or give it as one's reason with only a little thought. This kind of awareness of one's actions and motivations is all that traditional conceptions of action require, suggesting that habitual action can straightforwardly be classed as agential by traditional conceptions of agency.

In contrast, the influence of situational control on action is not the kind of thing that can be considered an instance of agency according to traditional conceptions. When an agent acts as the result of a situational cue she is unaware of, in a way she is unaware of, and in a way that cannot be rationalised by her explicit attitudes or the reasons she takes herself to have, then there seems to be no good reason to attribute the action to her as a person. For instance, if a committed egalitarian sits further away in an interview with a black candidate than a white candidate because of an implicit bias she rejects, without knowing that her behaviour is being influenced by a bias and without knowing that she is in this way discriminating, there seems to be no good reason to class her behaviour as an instance of agency rather than mere behaviour.[8] Though her body and her cognition partly caused her to sit further away, this kind of role in the act's causal history does not seem to be the right kind of role to say that what was done was an instance of the agency distinctive of persons, any more than would be said of a movement in sleep caused by the cognitive processes of a dream. The behaviour is not known under any description, nor is it performed due to a belief that it will satisfy some desire. Further, it is performed regardless of the reasons that the agent takes there to be for action.

However, the influence of autonomous goal control cannot be accounted for so simply or satisfactorily by traditional conceptions of agency. Traditional conceptions of agency exclude the behaviour guided by autonomous goal control from agency since it occurs beyond awareness and direct deliberative control. Unlike

[8] There may of course be a separate question concerning whether she is still morally responsible for her bias and the behaviours it encourages. This question should, however, be treated separately from the question of agency.

habitual action, behaviour governed by autonomous goal control is often not known under any description and cannot be intervened in directly by snapping back to full conscious control. However, it is not entirely beyond the agent's control. It is controlled by the agent in a distinctive way: what autonomous goal control she has is shaped by the goals she consciously pursues over time – often the things that matter enough for her to persistently strive to achieve them. Thus, both the way that autonomous goal control is established and the ends that it serves are integrally related to the person's own values and explicit attitudes, such as her beliefs about how she ought to act. For this reason, behaviour governed by autonomous goal control does seem like an instance of the agency distinctive of persons.

As can be seen from the above, behaviour governed by autonomous goal control prompts conflicting intuitions about whether or not it is an instance of agency. On the one hand, behaviour that an agent does not know about and cannot bring under direct control does not seem agential. On the other hand, behaviours that are ruled by and expressive of the things that most matter to agents as persons do. Two ways of resolving the conflict present themselves. First, one could give an account of agency that provides a principled and convincing way to accommodate or reject the full range of implicitly-governed behaviours, including autonomous goal control. Alternatively, one could show that one or both of the conflicting intuitions creating the tension can be safely ignored. What follows here is an attempt at each strategy. The intuition that behaviour arising from an agent's values is agential, even if other typical features of agency are lacking, is explored by examining John Doris's valuational model of agency (2015). Next, a deflationary response suggested by the work of Joshua Knobe (2006) will be considered, according to which intuitions about problematic cases of agency are dismissed as evidence about what kinds of acts there really are. Ultimately, neither versions of these responses considered here are found to be convincing, suggesting that pursuing a better way to resolve the conflict may be a fruitful line of philosophical inquiry.

## Chapter Two: Doris's Valuational Model

In *Talking to Our Selves,* John Doris sets out to achieve two goals: to undermine what he calls 'reflectivism' about morally responsible agency and to replace it with a 'valuational' model (2015, p. x). Reflectivism, according to Doris, is the doctrine that morally responsible agency is intentional action which is 'ordered by self-conscious reflection about what to think and do', and that its successful exercise requires 'accurate reflection' on salient aspects of the self, such as one's reasons for action (2015, p. 19). To undermine this doctrine Doris presents a sceptical challenge founded on empirical evidence of non-rational, non-conscious influences on intentional action (2015, pp. 41-77). Doris argues that reflectivists cannot rule out the possibility that any given action is caused by non-rational, non-conscious influences rather than by accurate reflection and that, if reflectivism is true, this entails scepticism about morally responsible agency.

Rather than concede to the sceptics, Doris aims to replace reflectivism with an account of morally responsible agency that is both compatible with the empirical evidence and that justifies typical attributions of moral responsibility. Doris's account is characterised by two central claims. First of all, it is what Doris calls a 'valuational account' since it defines morally responsible agency as those acts which express the agent's values, whether or not her actions are reflectively guided and whether or not she is aware of the values expressed (2015, pp. 24-33). Secondly, in a move away from the individualist picture of agency, Doris presents agency as a social and not an individual achievement. He argues that the processes enabling agents to discover and act upon their values are social, collaborative processes – for Doris, 'agents are negotiations' (2015, p. 148).

Although Doris is a pluralist about agency, he places the greatest emphasis on the importance of value-expression (2015, pp. 171-177). His hunch is that it is the fact that an act is aligned with and partly caused by the agent's values that really matters for agency. As such, his account might provide a principled and convincing way to differentiate between different kinds of implicitly-governed acts. If it implies that autonomous goal control and habitual action (but not mere situational control) are instances of agency then it might provide an intuitively plausible way to accommodate the influence of implicit cognition within a theory of action. Furthermore, since Doris's account is motivated in part by psychological evidence of some kinds of implicit cognition, it uses an interdisciplinary approach uniquely appropriate to the task in hand. However, it will here be maintained that some aspects of both Doris's method and his model ought to be rejected. First, Doris's method of holding agency and responsibility so close together leads to confusion rather than clarity. Second, closer consideration of the non-conscious influences that Doris claims undermine agency reveals that many of them pose no real threat to agency, which at most provides a way to defang the sceptical challenge and at least warrants more careful disambiguation between heterogeneous examples. Third, Doris arguably provides insufficient reason to think that it is an agent's values that matter, either for agency or responsibility.

## *2.1 Reflectivism and the Sceptical Challenge*

The first aim of Doris's project is to undermine a doctrine in moral psychology that he terms 'reflectivism': a doctrine about morally responsible agency that defines it by its relation to agents' accurate, self-conscious reflection. Before considering Doris's argument in detail, it is worth highlighting the notion of agency that he invokes. Doris is concerned with what might be described as a 'thick' notion of agency that is bound up with moral responsibility. Morally responsible agency is, for Doris, the kind of intentional action that is 'self-directed' (in a manner fleshed out below) in such a way as to make the action apt for moral evaluation, wherever the act is morally relevant (2015, pp. 23-25). To use his example: it is the agency that the young man Smitty exercises when he chews one's slippers as a prank rather than the kind of agency exercised by a puppy that chews one's slippers out of puppyish exploration (2015, p. 23). This sense of agency is, accordingly, not the 'thin' psychological sense of agency in play when describing the mere capacity to initiate one's own bodily movements, such as the kind of agency shared with animals. Nor is it reducible to the notion of intentional action: for Doris, intentional actions performed under duress or due to compulsive addiction are not instances of morally responsible agency, since they are not self-directed in the right way (2015, p. 25). For ease, in this chapter 'agency' will be used mostly to refer to Doris's understanding of it, unless otherwise indicated.

Doris defines the reflectivist account of morally responsible agency as follows: reflectivism is the doctrine that 'the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do', along with the corollary claim that 'the exercise of human agency requires accurate reflection' (2015, p. 19). Put more loosely, agency on this model has to do with being able to reflectively discern things about oneself, such as one's desires and reasons, and act accordingly upon them. A typical instance of agency, according to reflectivism, would be something like this: Samira wonders what she would like to eat for lunch, decides she will have pasta because she likes it and it will provide energy for her run, and so she eats the pasta. Further, the reflectivist maintains that her acts of choosing and eating are examples of agency not only because they appeared to her to be governed by her reflection but also because the reflection was accurate – she does indeed like pasta and does plan on going for a run, and these were indeed the reasons that she chose and ate the pasta.

To undermine reflectivism, Doris draws on a range of studies in psychology that reveal non-rational implicit influences on human behaviour. He argues that if what these studies purport to show is true then reflectivists are forced into scepticism about morally responsible agency, since they cannot confidently rule out the possibility that a piece of apparently reflective action was actually caused by non-rational implicit cognition instead. The examples vary considerably but, according to Doris's analysis, they are all cases of a particular kind of 'incongruence' – the psychological mismatch that occurs whenever 'different psychological processes issue divergent outputs regarding the same object' (2015, p. 51).

The more innocent and familiar kinds of incongruence include *akrasia*, when an agent desires to do one act but judges that she ought to do another, and known perceptual illusion, when an agent perceives an object as being in one state but judges it to be in a different state. Acts influenced by these kinds of incongruence typically engage the agent's reason in some way: the akratic agent typically reasons about what to do but gives in to her desire, while an agent who acts upon her beliefs rather than her illusory perceptions obeys the output of her reasoning. The same cannot be said, according to Doris, of the more troublesome kind of incongruence that leads scepticism – what Doris calls 'incongruence with bypassing' (2015, p. 52). Bypassing occurs when an agent's action or judgement is caused by implicit cognition which, according to Doris, does not engage with reason at all – it merely bypasses it, such as when an agent whips her hand away from a hot stove. When bypassing and incongruence combine, often the result is that an agent's action is caused by cognition of which she is unaware and which she would not recognise as a justifying reason for her action were she to become aware of it. When this occurs, Doris characterises the cause of behaviour as a 'defeater' of agency: since the agent has acted due to a cause she would not recognise as a reason, were she aware of it, her behaviour is not an instance of morally responsible agency (2015, pp. 64-65). Doris discusses a wide range of examples of defeaters. Three of his examples are used below to illustrate the account he gives of the phenomena.

One example of a defeater Doris discusses is the 'Watching Eyes Effect', revealed by studies that suggest people are more likely to behave cooperatively if pictures of eyes are displayed within view than if images of other things, such as flowers, are displayed in the same place (2015, p. 41). For instance, studies found that participants are more likely to contribute to the coffee-money tin in their office (Haley & Fessler, 2005), to be more generous in giving money to shared resources in computer games (Burnham & Hare, 2007; Haley & Fessler, 2005), and to be more likely to tidy up after themselves in a shared cafeteria (Ernest-Jones, et al., 2011) when pictures of eyes were present. Doris maintains that the agents whose actions were influenced by the eyes are unlikely to recognise a picture of eyes as a reason to give money, share resources, or tidy up – instead, their actions were caused by defeaters and were thus not instances of morally responsible agency.

Another example is the ‘Ballot Order Effect’, which suggests that in some elections people are more likely to choose the name at the top of the ballot (Webber, et al., 2014; Meredith & Salant, 2013; Marcinkiewicz, 2014 & Krosnick, et al., 2004). These studies are performed by analysing patterns in real voting data, rather than conducting controlled experiments. One benefit of this is that it provides potential evidence of how implicit cognition affects behaviour in everyday settings. However, it does make it more difficult to determine what kind of cognition influenced agents’ decisions since they cannot be consulted or artificially constrained to certain types of thinking. Doris’s analysis runs as follows. He suggests that at least some voters must be influenced by the ballot order for there to be a statistically significant effect. He then argues that these voters presumably wouldn’t consider being on the top of the ballot a reason for voting for someone. If this is the case, he suggests that their action is likely to have been caused by a non-rational implicit cognition that they would not recognise as a reason (2015, p. 56).

Finally, Doris cites studies of implicit bias, studies that reveal that many agents harbour biases about groups such as black people, women and people with disability that can affect their behaviour implicitly, even causing them to act against explicit egalitarian values. Doris cites the study described above (Bargh, et al., 1996), in which non-African American participants had to complete an onerous task on a computer. In between tasks, pictures of either black faces or white faces flashed on the screen for times too short for participants to notice. When the computer appeared to crash and the participants were told they had to start again, researchers measured the hostility shown by participants. They found that participants shown faces of black people were more hostile than the control group, and that this effect was not significantly linked to self-report of racist or non-racist attitudes. These findings suggest to Doris that those participants committed to non-racist values would not recognise the influence of black faces as a justifying reason for their increased hostility, so that a defeater obtained (2015, p. 57).

Having set the scene, Doris’s sceptical argument is fairly easy to characterise. He argues (pp. 64-65):

> Where the causes of her cognition or behavior would not be recognized by the actor as reasons for that cognition or behavior, were she aware of these causes at the time of performance, these causes are defeaters. Where defeaters obtain, the exercise of [morally responsible] agency does not obtain. If the presence of defeaters cannot be confidently ruled out for a particular behavior, it is not

> justified to attribute the actor an exercise of agency. If there is general difficulty in ruling out defeaters, skepticism about agency ensues.

In a nutshell, Doris argues that if reflectivism is true then there is going to be a general difficulty ruling out defeaters. To see why this might be the case, consider Samira's decision to eat pasta. In order to rule out defeaters, the reflectivist would need to show that there is good enough evidence to believe that Samira acted according to her reflective reasoning and not because of some cause that she would not consider a reason, such as having walked past a pasta advert that she didn't notice. Given that the psychological record suggests such influences occur frequently outside of the agent's awareness, Doris maintains that it is a live hypothesis that any experience of reasoning or deciding that Samira might present as evidence of a reflective act might have been influenced by a defeater without her knowing it.

The reflectivist might respond that the psychological evidence is insufficient in the face of substantial everyday experience of reflective agency. She might argue that we choose to do things all the time after thinking about what to do, and that the burden of proof lies with the sceptic to show that anything is amiss in these everyday cases. In response, Doris attempts to undermine the evidential value of the experience of agency. He does this by outlining a wider range of studies that reveal self-ignorance, arguing that they (as well as the cases considered above) suggest that agents often do not have an accurate understanding of what guides their actions. If they do not, Doris argues, then neither their experience nor their report of reflective agency can provide sufficient evidence of reflective control. The sceptical challenge still remains: if we are often mistaken about the cognitions that cause and explain our actions then the possibility of mistake must be ruled out before an account of our own agency can be trusted.

For instance, Doris cites studies of the 'Group Effect', according to which the likelihood that an agent intervenes to help someone decreases according to how many other bystanders there are (Latané & Darley, 1970; Latané & Nida, 1981). Typically, agents do not report the number of other people around as a reason for not helping and deny that this was the cause of their decision when presented with experimental results suggesting otherwise, suggesting ignorance of the factors affecting their decision-making and action. Further, Doris discusses studies of 'choice-blindness', in which agents are asked to choose between a range of different options and then explain the reason for their choices. For instance, Johansson et al. asked participants to choose the most attractive face from a range of faces (2005), while Hall et al. asked participants to choose what attitude best reflects their own attitude towards a range of moral problems (2012). When asked to justify a choice different from the choice they have just made, the majority of participants fail to notice and give reasons they seem to believe in to justify a choice that they did not make. Finally, Doris discusses motivated forms of self-ignorance, according to which agents inaccurately judge themselves to be in some way better than they are – for instance, thinking themselves more talented (Cross, 1977; Zenger, 1992), more popular (Zuckerman & Jost, 2001), or more likely to stay married than they really are (Fowers, et al., 2001). Given that these facts about oneself will often be relevant to what one does, these inaccuracies are likely to be practically relevant.

If these varying studies do reveal that agents are commonly mistaken about the causes of their actions and the aspects of themselves relevant to their actions then, Doris argues, appealing to the experience of agency cannot successfully rule out defeaters. He summarises the problem thus (2015, p. 96):

> For experience to have epistemic heft sufficient to block skepticism, its reliability must not be subject to substantial doubts, but the empirical literature indicates that it is so subject.

In other words, if agents are often mistaken about the aspects of themselves that cause and explain their actions, we cannot appeal to their understanding of their own actions in order to rule out the possibility that that understanding is mistaken.

We can sum up the sceptical argument as follows: a wide range of studies in empirical psychology suggest that human actions are often caused by influences that agents are not aware of and would not recognise as reasons for their actions if they were to become aware of them. Moreover, there is evidence to suggest that agents are often mistaken about themselves and what causes them to act. Given these facts, it is a live hypothesis that defeaters might obtain for any instance of agency and our theory of agency ought to be one that can confidently rule out defeaters. Since reflectivism relies on the agent's accurate self-awareness – that which is put into question by the empirical evidence – reflectivism cannot confidently rule out defeaters. Accordingly, if reflectivism is true, we are driven to scepticism about agency.

## *2.2 A Valuational Account*

Having characterised reflectivism and the evidence from empirical psychology as outlined above, Doris faces two options. The first is to accept scepticism about morally responsible agency, while the second is to reject reflectivism and provide an alternative account that can rule out defeaters. Doris opts for the latter route and presents an account of agency that is '*anti-reflectivist*', '*valuational*', '*collaborativist*' and '*pluralist*' (2015, p. x, original emphasis). The central features of Doris's positive account will be briefly explored here before presenting criticisms of the picture as a whole. Though there are doubtless valuable things to be gained from studying the whole account in detail, the primary focus in this paper will be the model's core claim that what is needed is a 'valuational' account of agency – the idea that morally responsible agency is that which expresses an agent's values.

First and foremost, Doris argues that what is distinctive of morally responsible agency is that it is *self-directed* in a certain way – that is, governed by 'features of the self, such as desires or beliefs, as opposed to features of the environment that are "external" to the self, such as political regimes or natural disasters' (2015, p. 24). This seems a promising beginning but it raises the question of what should count as internal to the self and what external. Rather than engage at length with this notorious problem, Doris asserts: 'one has to start somewhere, and I'm going to blunder ahead with the notion of value, and say that behaviour is self-directed when it expresses the [agent's] values' (2015, p 25). The main attraction of this position, according to Doris, is the way that it aligns with moral evaluability. In a paragraph that is worth quoting at length, Doris seeks to illustrate this point (2015, p. 25):

> [When he chews my slippers] I'm not tempted to see my rowdy puppy as self-directed because I'm not tempted to see his behavior as an expression of his values (I don't think he *has* values), but I am tempted to see the deplorable Smitty's similarly destructive behavior as self-directed, since I suspect it expresses the value that Smitty places on a life of boisterous immaturity. The same expedient might be thought to distinguish different behaviors by the same [agent]: when the nicotine addict guiltily succumbs to craving and lights up, his behavior is not self-directed, but when he manages to resist a craving because he values his health, his behavior is self-directed.

According to Doris's analysis, what characteristically marks out a morally responsible act is that it is an expression of some value of the agent. Acts performed by agents who are supposed not to have values, such as puppies, are thus not instances of agency but mere behaviours, as are the actions of agents that are not expressive of their values – such as the act of an unwilling addict when she gives in to an overpowering craving despite not valuing the hit she thereby attains.

Doris's valuational account cannot be fully understood without considering what his notion of value-expressive acts amounts to. Doris first claims that: '[a] behavior expresses a value, we can say, when that behavior is guided by a value-relevant goal' (2015, p. 26). By this, Doris means that responsible agency is a subclass of goal-directed behaviour – what Doris understands intentional behaviour to be – where goal-directed behaviour is behaviour guided by a mental representation of some desired future state. Not all goals are value-relevant, however. Doris suggests this is because not all desires are equal. While some desires suggest that a person values the sought-after state, others are too fleeting, passing or alien to do so. He claims that 'values are associated with desires that exhibit some degree of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning' (2015, p. 28).

Though Doris never unpacks quite what is meant by 'associated with', he will here be read as suggesting the following: an agent values X when she has some desire(s) to bring about X that have, to some degree, the properties of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning. An agent's values, accordingly, are here understood as the things that she desires to bring about whenever her desires to do so have the relevant properties. For instance, Samira values having good health if she desires to bring it about that she has good health and her desire for it has the following properties to some extent. Her desire must have some significant degree of strength so that it cannot be outweighed by just any passing care. It must last some significant duration – as a standing desire if not always an occurrent one. The value (good health) must be desired for itself and not only as the means to an end, and it must be something that cannot be easily replaced by some other good. Finally, her desire must be the kind of desire that guides some of her behaviour, for instance by shaping a plan to achieve good health, and she must accept it as justifying her health-pursuing acts.

Contrasting pairs of examples can help clarify the difference between value-expressive acts and other intentional acts. Consider the cases of the willing and the unwilling addict. If both addicts badly desire a cigarette and then smoke one, both perform a goal-directed act according to Doris's account. However, only the willing addict performs a value-expressive act because of the difference between the addicts' desires. While the willing addict would take her desire to smoke a cigarette as justification for doing so, the unwilling addict would not. Alternatively, consider the difference between the reliable donor and the one-off donor. Let us stipulate that the reliable donor has durable, strong desires to help homeless people and so she gives to a local shelter every month. The one-off donor, on the other hand, experiences a fleeting, weak desire to help the homeless when she passes by a homeless young man and gives him some spare change. While the reliable donor performs a value-expressive act, the one-off donor does not since her desire lacks the strength and longevity required.

Doris adds a further clarification later in the text that values are expressed by actions only when the fact 'that the actor holds the value [is] causally implicated in her undertaking a behaviour suited to realize the value', with deviant causal chains ruled out (2015, pp. 135-136). For instance, Samira expresses that she values good health whenever the fact that she values good health causes (in the right way) an action of hers that furthers her chances of good health. If her value is part of what causes her to choose a healthy lunch, for instance, this act expresses her value and thus counts as an instance of morally responsible agency. If, however, she chooses a healthy lunch only because her workplace bans all unhealthy lunch options, or if she unintentionally faints at the prospect of an unhealthy lunch, then her value is not causally implicated in the right way and so no instance of morally responsible agency occurs.

Importantly, Doris maintains that an act can express an agent's values even if she is unaware of holding the value and/or unaware of how the act expresses a value she holds: 'people may have desires, values, and plans that they are quite unaware of, and their behavior may express their values without their knowing that it does so' (2015, pp. 27-28). This claim is necessary for Doris's account to stand apart from reflectivism and to account for agency in the face of pervasive self-ignorance. This is because allowing that an agent may express values she is unaware of holding allows her to be held responsible for acts despite being ignorant about the salient aspects of herself that brought them about.

### *2.3 A Social Account*

With this core element of his valuational account in place, Doris begins to paint an anti-reflectivist picture of how such value-expressive agency is typically brought about. Again drawing on psychology, Doris turns away from the individualist ideal of reflectivism towards a collaborativist, dialogic picture of agency. On Doris's model, agency is typically a 'socially embedded' achievement: agents typically and optimally reason, act, and discern salient things about themselves not through an individual effort to do so but as part of a negotiation with others (2015, p. 103). Though it is not the central focus of the criticism presented below, the social aspects of Doris's account will here be briefly sketched.

First, Doris argues that optimal human reasoning is typically socially embedded. He argues (2015, p. 115):

> [...] human beings typically reason best when reasoning takes the form of an ongoing social process. Sociality is not merely a *precondition* of optimal

> reasoning, like adequate sleep and nutrition; optimal reasoning is *itself* characteristically social.

To justify this claim, Doris draws together a range of findings suggesting that reasoning is often enabled and improved by social exchange. For instance, he cites psychological studies that suggest that working with others helps participants to more quickly identify errors in their reasoning (Hill, 1982) and to more rapidly spot rules that allow them to solve a set of problems than participants working alone (Schwartz, 1995). Further, he cites studies that suggest that agents are most likely to reason about their moral views when justifying themselves to others, which leads Doris to claim that 'moral reasoning is typically socially embedded reasoning' (2015, p. 119). While Doris proceeds cautiously, taking note of the ways in which groups can in some circumstances hinder rather than promote good reasoning, the general conclusion he reaches is that 'human problem solving is characteristically socially embedded, and this sociality facilitates optimal reasoning' (2015, p. 122).

Next, Doris extends his argument to agency itself. He maintains that exercises of agency are also often socially embedded in a way that helps his account overcome the scepticism that challenges reflectivism (2015, p. 129):

> The social exchange of explanations and justificatory narratives [concerning action] erects a scaffolding that supports behaviour expressing the actor's values. These dialogs effect the exercise of agency in conditions of self-ignorance where direct reflective control may falter. In the right social milieu agency obtains in spite of – or rather *because* of – self-ignorance.

Doris argues that agency is often enabled by the exchange of what he calls 'rationalizations': explanations of behaviour given by an agent or her interlocutors that make her actions intelligible, whether or not they accurately reflect the causes of her behaviour (2015, pp. 138-143). To return to an earlier example, such a rationalization might be Samira's claim that she ate pasta because she likes it and needs energy for a planned run. Doris suggests that such rationalizations, both provided by an agent's own understanding and by the accounts of others, form part of her 'biography' – the narrative by which she makes her behaviour intelligible (2015, pp. 143-146). Whether it is accurate or not, this biography can motivate the agent to act in ways consistent with it, helping her to express her values. For instance, Doris might say of Samira that some aspect of her biography, such as her self-understanding as someone who values being fit and healthy, might lead her to rationalize her choice of pasta by telling herself that she chose it to fuel her run. Both her self-understanding of being fit and healthy and this rationalization make it more likely that she will actually go on the planned run, since she wants to live up to her own and/or others' expectations. If she does then go for a run, one which expresses that she values keeping fit, then her biography has helped her to exercise her agency.

It may be objected that there need be nothing especially social about the picture Doris paints. One might invent one's own biography and strive to live up to it entirely independently of others' input. However, Doris convincingly argues that as we are highly social beings our biographies are in a large part 'negotiations' with others around us (2015, pp. 146-148). Agents are raised in families and societies in which certain roles are determined by shared understandings of what they entail, such as 'son', 'friend', or 'feminist' – roles that come with expected patterns of behaviour. Doris argues that agents adopt these roles through negotiation with others and use them to help structure their lives. They do so as agents – active 'participants' in their biographies – when they are able to accept or reject aspects of the roles they adopt and thus shape shared understandings of both who they are and the roles they perform (2015, p. 148).

In addition to this general picture of socially embedded agency, Doris provides examples of how particular social exchanges support agency. For instance, Doris presents the example of talking therapy (2015, p. 124):

> First, the "talking cure" is very much a social treatment, where client and therapist work things through together more effectively than the client could do on his own. There's the collaborativism. Second, decreasing psychological discomfort and increasing personal efficacy are very likely values many clients in therapy hold, so the clinical process is reasonably thought to facilitate the expression of these values. There's the agency.

Doris here argues that the patient is enabled to live her life according to her values because of the social process of working through challenges with the help of her therapist. This is one case of what Doris sees as a broader pattern in which social interaction helps us develop a narrative about our behaviour, one that helps structure and motivate our value-expressive acts.

Doris is keen to emphasize that this picture, unlike the reflectivist picture, does not require accurate self-knowledge; according to his account, agency can and often does rely on self-ignorance. To see how this may be true, consider Doris's examples concerning married couples (2015, pp. 131-133). According to a survey Doris cites (Fowers, et al., 2001), American married couples typically believe that the probability that their marriage will end in divorce is close to zero, despite it being common knowledge that around half of American marriages end in divorce. Doris argues that this motivated self-ignorance is also motivational: the illusion of bliss can help couples to do the work, such as attending couples' counselling, that makes it more likely that their marriage will last. In this case, Doris argues, ignorance about their own chance of divorce facilitates actions that allow couples to stay married, helping them to express that they value married life. Similarly, Doris presents the hypothetical case of a mediocre academic 'Professor Drudge' who believes that he is very talented (2015, p. 144). He suggests that Drudge's ignorance of his actual academic merit can motivate him to put the work in that is required to become very talented, helping him to express that he values being a talented academic. By emphasizing how self-ignorance can facilitate agency, Doris seeks to show how his account is preferable to reflectivism that requires accurate self-knowledge.

Having presented this overall picture of agency, Doris makes two refinements in order to ward off objections. First, he considers the objection that he has not thoroughly refuted reflectivism since there are plausible instances of agency that involve accurate reflective control without any dialogue or negotiation, as well as plausible instances of agency that are not expressions of any value. Doris's response to this objection is simple: he defines his account as being pluralistic about both agency and responsibility (2015, pp. 171-177). For instance, he accepts that different psychological processes, including both direct reflective control and dialogic self-direction, can lead to actions that express the agent's values. Further, he concedes that there may be some instances of agency without value-expression, such as when an agent acts according to reflective deliberation but not because of her values. While this response is permissible, it does seem in tension with the importance that he places on value-expression when outlining his account of agency. To what extent Doris's pluralism weakens his account will be explored further below.

The second main objection Doris considers is a new sceptical problem for his account: the problem of how to correctly determine whether or not an action expresses an agent's values. It might appear that all Doris has done is relocate the problem if it is just as hard to determine whether an action expresses an agent's values as it is to determine whether it was caused by reflective guidance (2015, p. 159). Doris concedes that correct attributions of value expression are difficult but that they do not to lead to scepticism. He maintains (2015, p. 164):

> Attribution of agency and responsibility may be warranted when a pattern of cognition, rationalization and behaviour emerges, and that pattern is best explained as involving the expression of some value. Whether a particular action expresses a value may be uncertain, but the emergence of trends across iterated cognitions and behaviour can underwrite confidence that the trend is to be accounted for by reference to a person's values, rather than a massively coincidental run of defeaters.

Accordingly, Doris argues that his account can rule out defeaters by locating an action in a pattern of thought and behaviour that is best explained by attributing some value to the agent. This is because it is more likely that a value explains the pattern than that many defeaters coincidentally created a pattern just like that of someone who holds the value. This has the consequence that while Doris adopts 'currentism' about responsible agency, insisting that all that determines whether an intentional action is an instance of morally responsible behaviour is whether it expresses the agent's values at the time it occurs, the process of attributing a value to an agent must take into account the history of the agent in order to identify patterns in her behaviour (2015, p. 30). For some actions, perhaps because no clear enough pattern has yet emerged that would support a confident attribution of value-expression, it may just be too difficult to confidently attribute agency. Doris alleges, however, that this epistemological challenge is less serious than the scepticism threatening reflectivist accounts. Because of the way that patterns of behaviour can help rule out defeaters, Doris maintains that his account generates less scepticism and uncertainty than reflectivism (2015, p. 164).

### *2.4 Values and Implicit Cognition*

As the reader will have noticed, there are distinct parallels between Doris's account and the discussion above concerning how to accommodate the influence of implicit cognition in our theory of agency. For instance, the acts governed by mere situational control were ruled non-agential, since an unknown reaction to an environmental stimuli that is not rationalised by the agent's explicit attitudes concerning what to do seemed like a poor candidate for agency. Arguing along similar lines, Doris rules out acts caused by defeaters since an action caused by something that the agent would not accept as justification fails Doris's counterfactual test for responsible agency. Furthermore, the above discussion suggests that autonomous goal control does bring about agency because of the way that the acts it governs align with and arise from long-held goals. Similarly, Doris argues that behaviour counts as agency if it expresses a value partly because valuing is conceived of as having longstanding desires that structure the agent's behaviour.

Admittedly, there are important differences too between the discussion above and Doris's account. While emphasis is here placed on alignment with explicitly held attitudes, Doris argues that an act can be value-expressive even when the agent does not know that they hold the value in question. Further, it is questionable whether Doris would consider acts governed by autonomous goal control as intentional acts, given the way that agents lack awareness of them. However, the similarities suggest that Doris's account (or one like it) might provide a more principled way to accommodate the influence of implicit cognition on action than the largely discursive, intuition-driven account provided above.

To see how Doris's account might accommodate the influence of implicit cognition, consider this brief sketch. Doris argues that agency is behaviour that is caused and guided by value-relevant desires. Acts governed by autonomous goal control arguably fall within this class of acts. It is highly likely that an explicitly-held goal that matters to an agent enough for her to strive for it repeatedly over time, automatizing pursuit of it to the extent that autonomous goal control is enabled, will involve desires that are longstanding, strong, ultimate, non-fungible, and play a determinative-justificatory role in behaviour. Though the operation of autonomous goal control does occur outside of the agent's awareness, making it distinctly unlike what is typically thought of as intentional action, nothing in Doris's definition of agency relies on an agent's awareness of what she is doing. Perhaps this is because it needn't. Perhaps Doris's account could be finessed further so that it not only excludes situational control but also includes those implicitly-governed acts that express an agent's values.

Unfortunately, however, Doris's account as it stands cannot provide a principled account of agency. As will be argued below, it fails to defend the central claim that value-expression is what matters for agency. Further, it has such counter-intuitive implications for more everyday instances of agency and responsibility that it seems unlikely to be an accurate account. Where this leaves the influence of implicit cognition on action is as yet unclear. It might be the case that a new account of agency, one that draws on the central intuition but does not have problematic implications, can more successfully accommodate the influence of implicit cognition on action. Alternatively, as will be explored in Chapter 3, it may be that the central intuition driving this investigation is itself suspect and ought to be ignored.

### *2.5 Objections*

Doris's account is appealing for a number of reasons. His comprehensive effort to draw on psychological findings in order to make philosophical arguments provides an excellent example of interdisciplinary work in moral psychology. The way that his account of agency includes some kinds of behaviour that are caused by non-conscious cognition correctly recognises that agency is not limited to directly deliberatively-controlled behaviour. Finally, Doris's attention to the social aspects of agency provides a welcome alternative to individualist picture that arguably misrepresents profoundly social beings. However, it will here be argued his account should not be accepted in its current form, for reasons that will be explored below.

Firstly, Doris's accounts runs agency and responsibility very close together into what he admits might be a 'mare's nest' conception of morally responsible agency (2015, p. 32). Holding the concepts of agency and responsibility so close together leads to more confusion than clarity. This can be seen by considering cases in which an agent acts intentionally but does not act upon her values. Doris discusses some examples of this kind and presents them as cases of intentional behaviour that are not yet cases of agency since they do not express an agent's values (2015, p. 25):

> […] there are many breakdowns of agency, like succumbing to coercion or addiction, that are perfectly good candidates for intentional action. It's not as though the robbery victim surrenders his wallet or the nicotine addict lights up by accident; these are things the actors mean to do, but they're not the kind of things that make promising candidates for agency.

Even if we remind ourselves throughout this passage that 'agency' must be read as 'morally responsible agency', the notion of intentional behaviour that is not an instance of agency still seems strained, forcing together separate questions of what makes for agency and what makes for responsibility. To put it bluntly: addictive and coerced acts still seem like acts – not happenings – and though Doris acknowledges this by acknowledging their intentionality his refusal to treat them as instances of agency lacks rigorous defence.

Doris argues that intentional acts that do not express the agent's values are intentional because they are goal-directed – they are 'guided by a representation of a desired future state' (2015, p. 26), such as the representation of relieving a craving or appeasing a threatening robber. However, he suggests that they are not instances of agency because the desires these actions serve are not associated with any value the agent holds. For instance, the unwilling nicotine addict acts on her desire to smoke a cigarette but does not value smoking the cigarette, according to Doris's definition of valuing. This might be, for instance, because the desired end of smoking a cigarette is fungible or because appeasing a violent robber is only desired as a means of escaping harm.

While the distinction between behaviours driven by these desires and behaviours driven by value-relevant desires is tolerably clear, what is not clear is that this distinction amounts to the difference between agency and non-agential behaviour. If Doris argues that agency is the product of self-direction, he owes us an explanation of why the desires that are not associated with any value do not count as part of the agent's self while those relevant to a value do. This is not an argument that Doris provides. The closest he gets to an argument for why only value-expressive acts count as self-directed is his argument that values matter for responsible agency because values are fitting targets for reactive attitudes like praise and blame. This argument only suggests that value-expression should be what grounds moral responsibility – it is silent on what should ground the self-direction characteristic of agency. By holding responsibility and responsible agency so closely together, Doris uses an argument concerning responsibility to defend an account of agency, confusing the issue rather than clarifying it.

Arguably, it would be clearer to maintain that all intentional actions are instances of agency but that only value-expressive acts can be acts that agents are morally responsible for, separating the question of agency and responsibility. However, even if Doris's distinction was understood in this way it would be unlikely to satisfy many of his opponents. In the cases of unwilling addictive behaviour and coerced behaviour, at least some of us have moral intuitions that align with Doris's; there is something right in the idea that such agents are not fully morally responsible for their acts due to the pressures of addictive desires and threats of violence. However, other cases of intentional acts that do not express the agent's values evoke opposite intuitions. For instance, consider one-off instances of acting against one's values. Some Scrooge who does not value beneficence might be generous to a poor child just once on a kind whim, or an uncharacteristically spiteful remark might be made by some Saint who values kindness. Such people would typically be considered morally responsible for their deeds, despite the fact that they do not express the agents' values. Doris might account for such acts in one of two different ways. They might be understood as cases in which the agent does not have the value their one-off act seems to express, perhaps because Scrooge's desire to help or Saint's desire to hurt lack the requisite strength or duration for their acts to count as value-expressive on Doris's terms. Alternatively, they might be understood as cases in which the actions do express a value but onlookers do not have sufficient grounds to attribute the value to the agent because no pattern has yet emerged that the value can explain. According to either understanding, agents could not be properly praised or blamed for such acts, a feature of Doris's account which many will find unconvincing.

In response to these objections, Doris would perhaps emphasize his claim that it is right to hold responsibility and agency close together because 'typically, instances of morally responsible behaviour involve exercises of agency' (2015, pp. 158-159). There are two ways in which this defence would fail. Firstly, it might be rejected due to the fact that people are often held responsible for such things as unintentional omissions, attitudes and emotions that do not seem to involve agency. Samuel Murray's commentary (2018), for instance, rejects Doris's model because of its inability to explain the widespread practice of holding agents responsible for unintentional omissions, such as forgetting to call one's friend to wish her happy birthday. Doris's response to the problem of unintentional omissions is to suggest that, in many cases, what agents are actually morally responsible for is not the omission itself but a prior morally responsible act that led to the omission, such as choosing not to write friends' birthdays in one's diary. When an unintentional omission is not the consequence of any previous responsible act, Doris maintains that we ought not hold the agent responsible precisely because it does not express her values or derive from any prior value-expressive act.

A full discussion of the correct account of responsibility for unintentional omissions is beyond the scope of this dissertation. However, it will suffice here to note two things. First, insisting that those who are responsible for omissions are really responsible for a prior intentional act can have implausible consequences, as Doris notes: the drunk-driver ceases to be responsible for hitting a pedestrian and becomes only responsible for drinking in circumstances in which she was likely to drive a vehicle (2015, p. 31). Second, even if Doris's account of responsibility for omissions is correct it may struggle to account for a wide range of other things that agents are often held responsible for that arguably do not involve or result from agency, such as one's attitudes, beliefs, emotions, values and character. It seems unlikely that those held responsible for excessive anger, for instance, are really responsible for some prior intentional act that led to their anger.

The second reason that this defence would fail that it is not at all clear that just because the objects of moral responsibility often involve an action that it is right to analyse agency in terms of moral responsibility. To see why, consider other things that are typically involved when the question of responsibility arises, such as the agent's capacity to reason or the possibility that she might have acted otherwise. It does not follow that because these features are typically involved in morally responsible acts that they themselves should be analysed in terms of moral responsibility. The capacity to reason, for instance, should not be defined as the capacity to reason about right and wrong simply because this capacity is invoked in judgements of moral responsibility. Similarly, it does not follow that agency should be understood as value-expression just because value-expression is often invoked in assessments of moral responsibility.

Seeking another way to defend his position, Doris might fall back on the pluralism of his account which allows that there may be agency and/or responsibility without value-expression (2015, pp. 171-177). While this is certainly a permissible and plausible response, it is problematically in tension with the central place of values in Doris's account. If there can be both agency and responsibility without value-expression, it becomes less clear why Doris insists that value-expression is of such central importance for morally responsible agency – especially if there is independent reason to question whether values really matter for agency, as will be explored below. The possibility of agency without value-expression makes it open to the critic to suggest that perhaps values do not really matter for agency. Perhaps, instead, some other feature shared by both reflectively-guided acts and value-expressive acts is what really matters. A candidate feature might be what Doris describes as self-direction. Doris attempts to argue that agency is behaviour characterised by self-direction, a feature that he then analyses as amounting to value-expression. However, if Doris concedes that there can be agency without value-expression it becomes an open question how self-direction should be understood, since presumably instances of agency without value-expression are self-directed in some other way. If self-direction could be analysed without presenting it as reflective guidance or as value-expression, in such a way that still warded off Doris's sceptical threat, Doris's account would appear explanatorily weak by comparison.

The second challenge here considered regards the sceptical argument which leads Doris to reject reflectivism – specifically the central claim that causes of behaviour that the agent would not recognise as providing justifying reasons for her behaviour count as 'defeaters' of agency. There is good reason to question the assumption that the causes of behaviour Doris discusses really defeat agency, even on his terms. To see why this assumption may be too quick, consider again the Watching Eyes Effect. The very fact that images of eyes – and not images of flowers – inspired greater generosity suggests that that the image is engaging with something like a good reason for generosity, such as an awareness that others depend upon your cooperation or that your reputation might suffer if you are caught being selfish. If the Watching Eyes Effect is in fact mediated by some value of the agent's, like her value of being cooperative or being well-liked, then its apparent threat to agency shrinks remarkably.

To emphasize the above point, Zina B. Ward and Edouard Machery (2018) compare the influence of defeaters with other rationally arbitrary features of the agent that influence what she does. Facts like where she grew up, what kind of personality she has and what her parents are like are all rationally arbitrary influences on an agent's behaviour. These influences are all such that she might not recognise them as providing justifying reasons for her actions. For instance, what her parents are like might influence her to prefer an environmentally-friendly lifestyle, but she may not see this as a good reason to protect the environment. However, if such parental influence is mediated by her values then it need not undermine her agency. If she lives an environmentally friendly lifestyle partly because of her parents' influence and partly because she values the environment then her agency seems secure, according to Doris's model. As for these factors of upbringing, Ward and Machery argue (2018), so for apparent defeaters.

Another way to undermine Doris's worry about defeaters is to consider how threatening Doris's defeaters are to agency in cases in which the agent is plausibly ambivalent about what action she performs. For instance, the Ballot Order Effect has been found to be most pronounced for two types of voter: those with little information about the candidates and those who are nearly indifferent to the choice that they make. As Neil Levy suggests (2018, p. 31), if the Ballot Order Effect influences those who either know too little or care too little to make a choice that expresses their values then it arguably does not undermine their agency, since their choice of the top-listed candidate will be no less value-expressive than a choice that was governed solely by deliberation. What prevents their value expression is not the ballot order effect, Levy suggests, but their ignorance and/or their lack of preference. Further, mediation might matter here too. While an agent might reject 'The name is at the top' as a good reason for voting for a candidate, they might accept 'The name is at the top and, since I do not really know who to choose, I might as well pick the first one rather than read on carefully and think about them all'.

Murray suggests that ambivalence might play a role in cases of choice blindness, too. He focuses on Hall's studies involving moral statements, drawing attention to the fact that the statements used were fairly complex and that the same results might not have obtained were simple statements such as 'Torture is wrong' used instead. Then he suggests (2018):

> Perhaps with complex moral statements about [things like] the permissibility of government surveillance, people are ambivalent. They can think of reasons that support mutually exclusive positions. The initial judgment accords with the reasons that the subject finds salient at that moment. When the experimenters reverse the answer [...], perhaps the switched answer makes salient the considerations in favour of the competing position.

If agents subject to choice blindness were ambivalent about the statements they read then either agreeing or disagreeing with them would be equally value-expressive, making the effect of their choice-blindness seem less threatening to agency on Doris's terms. If the agents are ambivalent, this might also suggest that other values will have a stronger rational hold on their behaviour, such as the value of saving face by giving convincing reasons for statements one is supposed to have chosen rather than seem ambivalent about an important issue. If agents were ambivalent about the statement but motivated by other values then choice-blindness seems less threatening to agency.

However, not all of Doris's examples can be so easily dismissed as posing no real threat to agency. For instance, it does not seem likely that the influence of an unwanted implicit bias is mediated by any value of the agent or other reasons that she has for acting – unwanted bias seems to influence egalitarians to act not only upon causes that they would not recognise as justifying reasons for their discriminatory behaviour but to act in ways that directly contravene values they hold. For this reason, it also seems likely that at least some egalitarians exhibiting inegalitarian implicit biases are not merely ambivalent about the influence of the bias – it is more likely they would strongly disapprove of it were they to become aware of it. Accordingly, the foregoing considerations do not suggest that there is no such thing as a defeater. However, they may suggest that the effects Doris terms 'defeaters' are a heterogeneous group and that he might improve his account by following Enoch Lambert and Daniel C. Dennett's recommendation that he recognise 'a gradient between goofy and not-so goofy' influences on behaviour (2018, p. 29).

Finally, several authors (such as Fowers et al (2018) and Lambert and Dennett (2018)) have called upon Doris to take individual differences more seriously when considering whether the non-conscious, non-rational influences on behaviour that he cites are defeaters of agency. The kind of experiments in social psychology that Doris draws his evidence from rarely study individual difference, making it difficult to paint an accurate picture of the extent to which vulnerability to Doris's defeaters varies between individuals or to understand why people differ in this respect. If, for instance, research into the influence of virtue and character on susceptibility to defeaters found that the possession of certain virtues or character traits had a significant influence on how likely an agent is to be susceptible to defeaters then the sceptical challenge may look less worrying. This might be because the cultivation of virtues and character traits is itself considered agential, so that cultivating a virtue that leads one to behave (or not behave) in a certain manner is a way of indirectly exercising one's agency. Alternatively, it might be argued that virtues, vices and character traits make proper objects for moral assessment, which may bring the influence of defeaters within the scope of moral responsibility if it was moderated by character. It would be especially damaging for Doris's argument if it could be shown that individual differences in susceptibility to defeaters are shaped primarily by prior reflective activity.

The final and most central reason to reject Doris's account – one that has been suggested by earlier criticisms – is that he fails to make a strong enough case that values are what really matter for morally responsible agency. According to Doris, morally responsible agency is intentional behaviour that is self-directed, where self-direction is analysed in terms of values: intentional behaviour counts as agential if it is self-directed and it counts as self-directed if it expresses a value held by the agent at the time of acting. However, there are a range of problem cases for this model of agency which, taken together, suggest that values cannot provide the justification Doris seeks. These problem cases suggest either that something different or that something more is needed to ground morally responsible agency, rejecting the heart of Doris's account.

This objection is often posed as part of a criticism of Doris's currentism. Doris insists that what typically determines whether intentional behaviour counts as morally responsible agency is whether it expresses a value held by the agent at the time of acting, regardless of how the value was formed. Critics such as Manuel R. Vargas object that this is not the case; Vargas argues that the history of a value matters since 'in the real world, [values] are too often the products of processes that are themselves culpability undermining' (2018, p. 49). Vargas's central example is the case of an agent who has developed adaptive preferences in oppressive circumstances so that, due to her restricted choice, she has adapted to prefer things which are 'counter to [her] flourishing or otherwise not what [she] would prefer under more normatively optimal circumstances' (2018, p. 49). Doris's model suggests that actions according to these values are self-directed in just the same responsibility-grounding way as those of luckier agents in situations who may choose preferences better suited to their flourishing. For critics such as Vargas, however, this is a counter-intuitive result. Vargas argues that such agents are not exercising morally responsible agency since they have not developed their values in conditions that really allow them to shape their own values, so that the very values thought to ground their morally responsible agency are not things that they are responsible for. To put this objection in Doris's terms: Vargas suggests that such agents are mere subjects in the negotiation that shapes their behaviour – not 'participants' – and the way to recognise this is to insist that only values formed in conditions that allow genuine negotiation can ground morally responsible agency.

A more fundamental way to challenge Doris's valuational model, one that has motivated some of the objections given above, is to simply ask what it is about values that makes for agency. Doris arguably provides insufficient reason to think that value-expression is what matters for morally responsible agency. For instance, as has been suggested, one might wonder why an intentional action that serves a long-standing, ultimate desire is more self-directed than an intentional action that serves only a passing fancy. As was suggested earlier, both desires are presumably internal to the self and many would argue that both are appropriate targets for moral praise and blame. Conversely, one might wonder why unintentional behaviour that expresses values is not an instance of agency if values are what count. Consider the case of the sleepwalker who commits a crime she would not commit while awake that nonetheless does express her values. Since Doris's account is limited to intentional actions, such a crime would be excluded from morally responsible agency but one might wonder why if it is value-expression that characterises self-directed, morally responsible behaviour.

Moreover, an opponent might argue that the difficulty of identifying what is agency-making about values is only increased by the fact that our values are shaped by so much external to our selves, such as the contingencies of our upbringing and the influence of those around us. We cannot see values as the mark of self-direction, it might be said, given that they are shaped by the social environment in which we find ourselves. Doris makes considerable allowance for this concern by building social influence into his very notion of self-direction: for Doris, the process of developing and understanding our values typically is social, a negotiation between the agent and others in her life. Doris suggests that the influence of others merely 'constrains' agency, rather than undermining it, whenever the negotiation is one in which the agent is a participant able to contribute in some way. The critics might insist, however, that Doris's accommodation merely highlights their challenges: they might argue that the importance of the agent's participation in the shaping of her values suggests once again that values alone are not what matters for agency – rather, it is whatever capacities or conditions allow her to participate in the negotiation of her agency rather than be merely subject to it.

John Doris is right to draw our attention to influences on our behaviour that fall beyond the bounds of reflective control and to ask questions concerning what the best picture of morally responsible agency ought to be, given that we are subject to such influences. Further, Doris is right to insist that our social nature be attended to when conceptualising agency and responsibility. However, he is too quick to interpret examples of bypassing with incongruence as undermining agency. Closer attention to the way that so-called 'defeaters' shape behaviour suggests that the influence of many of them does not pose so great a threat to agency as Doris suggests. Furthermore, the valuational model that Doris proposes does not seem up to the task of grounding morally responsible agency. It does not seem that value expression alone is what characterises self-direction or what grounds moral responsibility. Accordingly, we must continue investigating non-conscious influences on human behaviour, seeking an account of agency and moral responsibility that can accommodate them more successfully while still providing a plausible account of what makes for a morally responsible act.

## Chapter Three: Intuitions and Agency

So far, this investigation has proceeded on the assumption that our intuitions often provide useful evidence about agency. The method adopted throughout has been to treat intuitive judgements about instances of behaviour as starting points for philosophical investigation into what kinds of action there are – an instance of what Matthew Liao has termed the 'IAE' or 'Intuitions as Evidence approach' to philosophy, whereby intuitions and the conflicts between them are treated as useful evidence for philosophical research (2008, p. 248). For example, considering the different ways that implicit cognition influences behaviour in Chapter One brought out contrasting intuitions that were taken as evidence that behaviour governed by autonomous goal control is agential while behaviour governed by mere situational control is not. The idea that values might matter for agency, arising from reflection on these contrasting intuitions, led to consideration of Doris's valuational model in Chapter Two. Intuitions about the cases Doris presents underpinned some of the central criticism made of his account, such as the intuition that intentionally performed acts are instances of responsible agency even if they do not express an agent's values.

The usefulness of the IAE method to this inquiry relies on the assumption that intuitions about agency often track relevant features of the behaviour considered; only if this is the case can intuitive judgements provide reliable evidence for inquiry. However, it might be objected that there is good reason to suspect this assumption.

In a series of papers in experimental philosophy,[9] Joshua Knobe presents evidence that some of our intuitive judgements of intentional agency are influenced by normative considerations that seem irrelevant to the question of what features a behaviour must have in order to count as an intentional action. If his analysis is correct then intuitions about agency may require a much healthier dose of suspicion than they have been treated with thus far. His analysis suggests a deflationary response to the central argument of Chapter One, suggesting that the intuitive judgements on which it rests may be explained away as spurious, morally-driven ascriptions of agency that, while they reveal a lot about our moral practices and concepts, are ill-suited to inform us about agency, and about what kinds of acts there actually are.

## *3.1 Joshua Knobe and Asymmetrical Intentionality Judgements*

In a range of experiments, Joshua Knobe claims to have revealed a surprising fact about folk psychology: intuitive judgements of whether an act is intentional (henceforth 'intentionality judgements') appear to be influenced by moral considerations in some 'intermediate' cases of intentionality (Pettit & Knobe, 2009, p. 587). By 'intermediate case', Knobe means an action that is neither paradigmatically intentional nor paradigmatically unintentional due to the fact that it has some features characteristic of intentional action but lacks others. For instance, consider an agent who does what she intends to do but only through chance since she lacks the skill to perform the act reliably. This would be an intermediate case since the act was performed according to an intention but it was not properly controlled by the agent's skill. Due to their intermediate status, intuitive judgements differ about whether actions in intermediate cases are performed intentionally. What Knobe seeks to demonstrate is that one of the features that can tip the balance when judging intermediate cases is the moral goodness or badness of the act.

The experimental case that has become most familiar is the case about a company chairman (or 'the chairman case'). In the original study (Knobe, 2003a), participants were presented with one of two vignettes describing a company chairman adopting a policy that increases profits but either helps or harms the environment. Participants were then asked whether the chairman intentionally brought about the

[9] See Knobe 2003a, 2003b, 2006 and Pettit and Knobe, 2009.

impact on the environment described, and also how much praise or blame the chairman deserved. The vignettes read as follows (2003a; 2006):

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

Before discussing the results, two features of the vignettes should be noted. First of all, both vignettes present intermediate cases of intentional action since they involve some but not all of its paradigmatic characteristics. Both are cases of what Knobe defines as 'side-effects': outcomes that the agent is 'not specifically trying to bring about' but that she foresees will be brought about by what she is specifically trying to do (2006, p. 208). Secondly, the actions are presented as being as similar as possible in those respects typically thought relevant to intentionality judgements, such as foreknowledge, choice and control, since Knobe is trying to suggest that they differ only in the goodness or the badness of their outcomes.[10] For instance, both chairmen report not caring about the environmental impact and explicitly state only being

[10] Knobe here relies on a consequentialist conception of morality. Since the chairman has the same motive in both vignettes, some might argue that the moral value of both acts is the same. One way to challenge Knobe, accordingly, would be to suggest that moral considerations cannot explain participants' asymmetrical judgements if those judgements were not made using a consequentialist framework. However, Knobe's consequentialist reading has here been accepted. This is done partly to provide a concise account and partly because Knobe's analysis can be challenged even without challenging his moral framework. If Knobe's analysis can be challenged even on his own moral terms, we perhaps have even more reason to reject it.

motivated by the profit. As such, these cases appear well designed to test whether moral considerations influence intentionality judgements in intermediate cases.

The first result of the initial study was that participants typically said that the chairman deserves a lot of blame in the harm condition but little to no praise in the help condition (2003a, p. 193). The second result is that there was a correlation between these judgements and intentionality judgements: participants were significantly more likely to judge that the chairman intentionally harmed the environment than that he intentionally helped it. In the initial study, 82% of participants judged that the chairman harmed the environment intentionally whilst only 23% judged that he helped the environment intentionally (2003a, p. 192), an asymmetrical response that is replicated to a similar degree in a range of subsequent studies.[11] This correlation suggests to Knobe that the moral evaluation of the outcome of an action can influence folk intuitions about whether the action was intentionally performed.

If Knobe's interpretation is correct it would be surprising since it would undermine the widely accepted 'Unidirectional Thesis' – the thesis that intentionality judgements are made independently of moral evaluations (Sripada, 2010, p. 160). Instead, Knobe's explanation suggests the 'Bidirectional Thesis' may be more accurate – the thesis that intentionality judgements and moral evaluations influence one another (Sripada, 2010, p. 160). If it turns out that the Bidirectional Thesis best explains the facts, this would not only undermine common thinking about the relationship between intentionality judgements and moral evaluation, it would also cast doubt on the usefulness of intuitive intentionality judgements when determining what kind of acts there are. The worry is well expressed by Chandra Sekhar Sripada as follows (2010, p. 161):

> It makes sense that descriptive features of a situation are processed prior to and/or independently of evaluative features, since to do otherwise would be tantamount to allowing one's evaluative reaction about what ought to be the case to inappropriately influence one's descriptive attitudes regarding what is the case.

[11] For example, see Nichols and Ulatowski, 2007; Phelan and Sarkissian, 2008; Sripada and Konrath, 2011. For discussion of studies that reveal the same effect using different methods and across different cultures and age groups, see Knobe, 2006, pp. 207 and 211-212.

In a nutshell, the worry can be understood as this: judgements that may be influenced by normative considerations seem ill-suited to the task of answering the descriptive question of what agency is like.

Knobe's original analysis of this finding (one he has since replaced) is useful to consider as it has shaped much of the subsequent debate. Rather than thinking that the asymmetry reveals a common mistake or bias, Knobe argues that it reveals the true nature of our competence with the concept of intentionality (2006, p. 204). He suggests the asymmetry reveals that the concept of intentionality is not just a tool for predicting, explaining and controlling behaviour but also a tool for assigning praise and blame, and that it has been shaped to suit this purpose (2006, pp. 227-228). Since different features of intentionality are relevant to the assignments of praise and blame in different cases, moral evaluation of a behaviour affects which features are necessary for judging that a behaviour is intentional, thus influencing intentionality judgements (2006, pp. 223-224).

Knobe's argument for the above conclusion proceeds as follows. First, he points out that in the chairman case and a range of similar studies there is 'a remarkable convergence between the conditions under which people assign praise and blame and the conditions under which they regard a behaviour as intentional' (2006, p. 224). From this he concludes that the concept of intentionality may permissibly be supposed to be a tool for determining how much praise or blame to assign to others (2006, p. 222). Next, Knobe notes that although each of the typical features of intentional action (such as trying, foresight and skill) play a role in the process of assigning blame and praise, they are of varying relevance to the moral assessment of different actions (2006, p. 223). For instance, the chairman's foreknowledge of the impact on the environment seems enough to make him blameworthy when he harms the environment though it is not enough to make him praiseworthy when he helps it. This suggests that there can be no way of grouping the relevant features of intentionality into a uniform concept able to help to determine praise and blame for all possible behaviours – any combination would be more relevant for some behaviours than others (2006, p. 223).

Because one uniform concept of intentionality would not be useful for the purpose of assigning praise and blame, Knobe argues that intentionality judgements are instead made using two sub-processes (2006, p. 226):

> The first sub-process takes in information about whether the behavior itself is good or bad and uses this information to determine which features are relevant. The second sub-process then checks to see whether the behavior in question

actually has these features and thereby generates an intentional action intuition.

Knobe uses this model to explain the asymmetry in the chairman case. He argues that participants first identify whether the outcome of the chairman's choice was good or bad, and that this influences their judgement of whether the chairman acted intentionally or unintentionally (2006, p. 226). When the outcome is judged to be bad, participants intuitively judge that either foresight or trying are sufficient for intentionality, leading to the judgement that the chairman acted intentionally. They then go on to consider how much blame to assign in light of this judgement. When the outcome is judged to be good, participants intuitively judge that both foresight and trying are necessary for intentionality, leading to the judgement that the chairman did not act intentionally. They then go on to consider how much praise to assign in light of this judgement. Accordingly, Knobe's model suggests that 'moral considerations are playing a helpful role in people's underlying competence [with the concept of intentionality] itself' (2006, p. 226).

Knobe's original analysis was met with a range of critical responses, most including experimental counter-evidence (for key examples see Phelan & Sarkissian, 2008 and Nichols & Ulatowski, 2007). In response to this criticism and after conducting further research, Dean Pettit and Knobe developed an alternative explanation that still gives moral considerations a core role in folk-psychological judgements (2009). In the experiments they discuss, participants were randomly assigned one or other of the chairman vignettes and asked to what extent they agreed with statements concerning the chairman's attitudes, with different statements being used to investigate the concepts of intending, desiring, deciding, advocating and opposing. For instance, the original question 'Did the chairman intentionally harm the environment?' was changed to the statement 'The chairman decided to help [harm] the environment' in order to investigate possible asymmetry in attributions of decision.

Pettit and Knobe argue that all of the concepts listed above are scalar in nature, so that judging that an agent holds them really involves picking out at what point the agent's attiude falls on a continuous pro-con scale. For instance, they claim that judging whether a chairman intentionally brings about X is really judging where his attitude towards bringing about X falls on a scale that ranges between something like 'is absolutely intent on causing X' and 'has no intention of ever causing X'. To capture this, participants were asked to indicate the degree to which they agreed with the statements about the agents' attitudes on a seven-point scale, where a score of 1 indicates strong disagreement while a score of 7 indicates strong agreement. Pettit and Knobe found a significant asymmetry between the help and harm case for every concept investigated (2009). While the asymmetry did not always result in judgements falling either side of the midpoint, in all cases there was a significant difference in degree of agreement between the help and harm case, with the chairman who harmed the environment being closer to the 'pro' end of the scale for all of the 'pro' attitudes and the opposite result obtaining when the concept investigated was the 'con' attitude of opposing. For instance, the mean attributions of desire for the harm vignette was a score of 3.4, while the mean score for the help vignette was only 1.6. While both of these averages suggest a general tendency to disagree, the asymmetry suggests to Pettit and Knobe that moral considerations have a significant influence on participants' intuitive judgements.

In order to explain this influence, Pettit and Knobe suggest the following framework (2009, p. 597):

> Pro-attitudes are assessed relative to a default, and this default is based in part on a sense of how things are supposed to be. The key claim then is that people's sense of what sort of attitude an agent is 'supposed to' have toward a given outcome can depend on the nature of the outcome itself. People are supposed to have more positive attitudes toward good outcomes, more negative attitudes toward bad ones. Hence, agents' attitudes toward these different outcomes end up getting compared to different defaults.

This model is easiest to grasp by way of an analogy that they provide (2009, p. 597). Suppose that someone were given a cup of coffee and a cup of beer that were both at room temperature and that she was asked to rate how warm each drink was. She might well respond that the coffee was cold and that the beer was warm – despite both being the same temperature – since she was judging their temperatures relative to the 'default' temperature that each drink should be served at. The authors suggest that a similar phenomenon explains the asymmetry in the cases studied. When an outcome is judged to be bad, people's attitudes towards it are judged relative to a default towards the 'con' side since one is not supposed to intend or desire such a thing (2009, pp. 597-598). A greater proportion of attitudes will fall on the pro side of a relatively con default, meaning that when an outcome is bad a greater range of midrange attitudes will seem like pro attitudes. When an outcome is judged to be good, people's attitudes towards it are judged relative to a default towards the 'pro' side since they are supposed to intend or desire such a thing. A greater proportion of attitudes will fall on the con side of a relatively pro default, meaning that when an outcome is good a greater range of the same midrange attitudes will seem like con attitudes. This model predicts the asymmetry Pettit and Knobe found.

It is beyond the scope of this thesis to arbitrate between Knobe's earlier and later analyses. What matters here is the fact that if either of his analyses are correct then we have good reason to think that intuitions about agency are influenced by moral considerations in intermediate cases. If this is true, it casts doubt on the value of such intuitions when investigating what kinds of actions there are. To see why this is the case – and particularly why it may cause problems for the current investigation – consider what it might reveal about the intuitions evoked in Chapter One. Consider two vignettes similar to the chairman case:

> Racist Bias
>
> A woman with explicit racist values also has an implicit racial bias against black people, one that has been strengthened by her past acts of wilful discrimination. In conversation with two salespeople in a shop, her implicit bias influences her to avoid eye contact with the black salesperson, discriminating against them in a way that is concordant with her values. Since her attention is fully taken up by her shopping and conversation, she does not directly deliberatively control this decision to avoid eye contact with the black member of staff and is not aware of doing it.
>
> Anti-Racist Bias
>
> A woman with explicit anti-racist values has an implicit racial bias against black people, one that is now often inhibited by implicit cognition that has been embedded by her past acts of wilful anti-racism. In conversation with two salespeople in a shop, her unwanted bias is prevented from influencing her behaviour and she makes eye contact with both members of staff. Since her attention is fully taken up by her shopping and conversation, she does not directly deliberatively control this decision to make equal eye contact with both members of staff and she is not aware of doing it.

The outcome of these cases – discrimination on the one hand and equal treatment on the other – seem likely to provoke a range of intentionality judgements. Many might deny that they are brought about intentionally since the agents are not actively trying to bring them about, however much they would endorse them if they were aware of them. Others might be more inclined to treat them as intentional due to the way they have arisen from the agents' values and past behaviours.

If Knobe is correct about the asymmetry of his findings, however, it seems likely that intuitions about the above cases will also be influenced by moral considerations. For instance, Knobe's analyses would both predict that folk would typically judge the former case as more intentional than the latter due to the immoral content and outcome of the racist agent's attitudes. While facts such as how much control an agent has over her behaviour and whether it aligns with her explicit attitudes might plausibly help to determine what kind of acts are presented by these bias cases, moral judgements about the goodness or badness of such acts seem unfit to underpin claims about what kinds of action they are. If the influence of moral considerations is pervasive, as Pettit and Knobe maintain, we have cause to treat intuitions about agency with a great deal more suspicion than they have yet met here.

## *3.2 Criticisms of Knobe's Analyses*

One way to defend the IAE approach would be to dismiss Knobe's analyses conclusively, thus undermining his claim that moral considerations influence intuitive intentionality judgements. This section presents a range of counter-evidence that, taken together, show that we can safely reject Knobe's explanations of his findings. First, experimental evidence is presented that Knobe's original analysis does not predict, and cannot satisfactorily explain. Experimenters have found that participants attribute intentionality in intermediate cases that are not judged to be morally or otherwise bad – indeed, it has been attributed in cases that are highly praiseworthy. Further, they have found cases in which a known, bad side effect is typically judged to be unintentional. Second, experimental evidence is presented that Knobe's later analyses cannot explain. Pettit and Knobe's own findings establish an asymmetry very like that of the chairman case when presenting vignettes that do not have a midrange pro-con attitude, a result they recognise their hypothesis does not predict. Further, statistical analysis of experiments conducted by Chandra Sekhar Sripada and Sara Konrath undermines the notion that moral considerations influence intentionality judgements.

Firstly, Knobe himself has found that side effects are sometimes judged as intentional when moral considerations are arguably irrelevant. In a study conducted with Gabriel Mendlow, Knobe adapted the chairman case so that the side effect was not helping/harming the environment but instead decreasing sales in New Jersey while increasing sales in Massachusetts to such an extent that overall sales increase (2004). Participants typically judged that decreasing sales in New Jersey was intentional but that it was neither blameworthy nor praiseworthy (2004, p. 257). Knobe's explanation for this result is that even though sales were increased overall participants judged decreasing sales in New Jersey as bad – though not morally bad – and that this badness made them more likely to deem the side-effect intentional (2004, p. 257). First, even if we accept Knobe's explanation of this result it is difficult to see how it can support the claim that specifically moral considerations influence intentionality judgements, since Knobe himself acknowledges that the perceived badness of this side effect is not moral badness. Second, critics Mark Phelan and Hagop Sarkissian present evidence that not even a judgement of non-moral badness explains the intentionality judgement in this case (2008). When they repeated Knobe and Mendlow's study, asking the additional question of whether decreasing sales in New Jersey was bad, only 14% of respondents claimed that it was – an insufficient proportion to explain the 64% claiming that the side-effect was intentional (2008, p. 295).

Perhaps the most powerful counter-evidence Phelan and Sarkissian provide, however, is the result of their city planner vignette study (2008, pp. 296-297). In this study they presented participants with a vignette in which a city planner decides to implement a programme that will clean up dangerous toxic waste in the city but also increase levels of joblessness. Though he reports feeling 'terrible' about increasing joblessness, he decides to adopt the programme in order to address the pressing problem of toxic waste. Though 69% of respondents judged that increasing joblessness was bad, only 29% judged that it was intentional. It is difficult to identify which feature (or features) explain the difference between the chairman case and the city planner case, since not only did the city planner express remorse for the bad effect but he was also motivated not by self-interest but by doing good for the community. Either of these features (and perhaps others) could explain the shift towards judging the side effect as unintentional. However, even without an alternative explanation the finding is enough to challenge Knobe's original analysis since it is a clear case of a bad side effect that is judged to be unintentional.

One final case arguably completes the challenge to Knobe's original analysis. As well as the chairman case, Knobe devised a sequence of vignettes designed to explore the influence of evaluative judgements on intentionality intuitions when the action is performed luckily – in other words, without the skill to perform it reliably. Intended acts that are achieved partly through chance are understood to be intermediate cases since, on the one hand, the agent has an intention to perform the act yet, on the other, she lacks the skill to reliably control her attempts. One of the vignettes in this study was that of Klaus, a selfless Nazi soldier who manages to shoot his own equipment in order to sabotage a mission and save innocent lives, even though he is 'not very good at using his rifle' and 'his hand slips on the barrel of the gun' making the shot go wild (2003, p. 320). 92% of participants judged that Klaus intentionally shot the equipment, despite lacking the skill to do so reliably (2003, p. 321). This provides an example of an intermediate case of intentionality in which the agent brings about a very good outcome and yet most people still judge the act as being intentional. Knobe's original analysis does not predict this result. If it were goodness and badness that made the difference then Klaus's good-but-fluky act should be judged unintentional, just like the chairman's good side effect. Taken with the other counter-examples presented here, the Klaus case suggests that the perceived goodness and badness of the outcomes are not what explains the asymmetry in the chairman case.

There is also evidence that counts against Knobe's more recent explanation. For instance, Pettit and Knobe managed to replicate the asymmetry typical of the chairman case in a pair of vignettes involving agents who are 'desperate' either to defuse or to detonate a bomb by guessing a pass code. Both are described as being fairly sure that they will not guess the code correctly. They then asked participants whether the agents 'intended' to defuse/detonate the bomb, and found that participants were significantly more likely to judge that the agent intended to detonate the bomb than that he intended to defuse it – results that parallel those of the chairman case. Pettit and Knobe's analysis does not predict that this combination of high pro attitude (being desperate to achieve one's ends) and low credence that one will succeed should be significantly influenced by moral considerations. This is firstly because both agents are described as desperately wanting to achieve their ends. Though one should be more inclined to defuse a bomb than detonate it, placing the 'default' position higher for the defusing vignette than the detonating vignette, this difference in default position should not lead to an absolute difference in intentionality judgements since both agents were described as being desperate to achieve their ends, putting their pro attitudes well outside the midrange. Though it is possible that low credence cases are also affected by moral considerations, Knobe and Pettit's analysis does not explain why. Since this result is not predicted by Pettit and Knobe's thesis, it suggests that it is not yet a correct (or at least a complete) explanation of the asymmetry.

Most convincing of all, Sripada and Konrath present findings that challenge both Knobe's old and new analyses – indeed, they challenge Knobe's central thesis that moral considerations influence folk intentionality judgements (Sripada & Konrath, 2011). These critics presented participants with one or other vignette from the original chairman case and then asked them six questions designed to measure a range of candidate explanations for the asymmetrical responses. Their intention in the study was to try to solve what they call the 'critical features' problem – the problem of how to determine which of a range of different candidate features really has a significant effect on some result (2011, p. 355). Participants were asked to indicate their views on the following topics by choosing a point on an appropriate 7-point scale (2011, see p. 356 for exact wording):

1. Whether the outcome was intentional (Strongly Disagree ... Strongly Agree)

2. How good/bad the outcome was (Very Good … Very Bad)
3. The moral status of the chairman (Very Moral … Very Immoral)
4. The chairman's environmental attitudes (Very Pro-Environment … Very Anti-Environment)
5. How likely the chairman was to bring about similar outcomes in other similar contexts (Very Likely … Very Unlikely)
6. Their own attitudes towards the environment (Very Pro-Environment … Very Anti-Environment)

The statistical analysis tool of structural path modelling was then used in order to establish, first, how much which case participants were given affected each of the candidate explanatory variables, and, second, how much each of these explanatory variables influenced the intentionality judgement made. What they found was that only the case given, judgements about the chairman's attitudes, and judgements about how likely the chairman was to act in similar ways in other contexts had a statistically significant influence on the intentionality judgement made (2011, pp. 361, 363). This means that none of the normative considerations (variables 2, 3 and 6) significantly influenced whether or not the side-effect was judged to be intentional, strongly suggesting that Knobe's central thesis does not correctly explain the asymmetry.

## *3.3 Alternative Explanations of the Asymmetry*

In the preceding section, evidence was presented suggesting that neither of Knobe's analyses successfully explains the asymmetrical responses to the chairman case. This tentatively suggests that he is too quick to conclude that moral considerations influence intentionality judgements. The case against Knobe would be strengthened, however, by a positive account of what really does explain the asymmetry. In the interests of thoroughness, a range of candidate explanations in the literature will here be briefly evaluated, before the most promising explanation is presented in the next section.

### Blame Bias

Philosophers such as Bertram Malle (2006) and Thomas Nadelhoffer (2006) have proposed that, instead of revealing the true nature of our core competency with the concept of intentional action, asymmetrical responses to the chairman case reveal a performance error due to a blame bias. Malle argues that when participants read the harm vignette they quickly judge that the chairman's actions are blameworthy. Their blame judgements then inappropriately bias them towards judging that the side effect was intentional, in order to provide post-hoc justification for their judgement of blame. Nadelhoffer's model is very similar, only instead of characterising blame as a judgement it is understood as an emotional, affective response to moral badness, following Mark Alicke's Culpable Control Model of blame (Alicke, 2000, cited in Nadelhoffer, 2006).

The bias account is fairly intuitive and conservative since it does not require radical revision of widely accepted accounts of our core competency with the concept of intentional action. However, it is arguably disproved by the cases examined above in which judgements of intentionality were made for side effects that were not perceived as blameworthy (Knobe and Mendlow, 2004; Phelan and Sarkissian, 2008). If the same intentionality judgements as are found in the harm case are found for side effects that are not blameworthy, this suggests that the blame bias model does not explain the asymmetry after all. Further, the blame bias account is undermined by the statistical analysis performed by Sripada and Konrath, who found that judgements of the chairman's moral status did not to have a significant influence on the intentionality judgements (2011).

### Pragmatic Effects

Frederick Adams and Annie Steadman also propose that the results of the chairman case do not reveal features of our core competency with the concept of intentional action. In fact, they argue that the folk do not have 'an articulated core [...] concept of intentional action' specifying the required mental states for attribution (2004, p. 176). Instead, they suggest that folk have an articulated concept of the pragmatics of intentional action and intentional action talk, one that has developed primarily because of intention's role in discussions pertaining to praise and blame (2004, p. 177). Their view is roughly that participants are reluctant to state that the side effect is unintentional in the harm case because stating that something was unintentional frequently implies that it is blameless, while stating that it is intentional typically strengthens the suggestion that it is blameworthy. Thus, because they want to blame the chairman, participants choose the answer that will have the pragmatic effects closest to the one they desire and state that he acted intentionally (2004, pp. 177-178).

Adams and Steadman's hypothesis predicts that if participants were given a way to express the chairman's lack of intention whilst also expressing disapproval and blame, participants would be more likely to class the harm as unintentional. Such an option would allow them to imply their blame while recognising that he was not trying to bring about harm. Unfortunately, when they tested this hypothesis their results arguably disconfirmed their thesis. They offered participants the choice between stating that the chairman 'knowingly and intentionally' harmed the environment and stating that he 'knowingly, but not intentionally' harmed it, hypothesising that the latter option would allow participants to imply blame as well as recognise his lack of intention (Steadman & Adams, 2007, pp. 27-28). Even given this choice, 80% of respondents still chose 'knowingly and intentionally' (2007, p. 28). These results replicate Knobe's findings even though participants were given a way to avoid unwanted implicature. Arguably, their results provide better support for Knobe, Malle or Nadelhoffer's explanations than the pragmatic account that they are trying to defend. In addition, when Nichols and Ulatowski performed a further experiment designed to test the pragmatic hypothesis (they gave participants the option to state that the chairman 'didn't intentionally harm the environment, and he is responsible for it' or to state that he 'intentionally harmed the environment, and he is responsible for it'), they once again simply replicated Knobe's findings (2007, pp. 352-353).

### The Trade-off Hypothesis

In an attempt to argue that moral considerations do not explain the asymmetry, Edouard Machery argues that it can instead be explained by what he calls the 'trade-off hypothesis' (2008, pp. 176-177). According to the trade-off hypothesis, folk typically judge that knowingly incurring a cost in order to obtain some desired benefit – a decision to 'trade-off' the cost for the benefit – is intentional. When presented with the harm and help case, participants judge that the harm case involves such a trade-off since the chairman accepts the cost of harming the environment for the benefit of increasing profits. Accordingly, participants judge that the harm is caused intentionally. However, helping the environment cannot be conceived as a cost since it is not negatively valued. Thus, they do not judge that a trade-off has been made and do not judge that it has been brought about intentionally. Machery's thesis avoids the revisionary implications of Knobe's explanations since it does not suggest that moral considerations influence intentionality judgements. However, there is reason to doubt that the evidence Machery advances for his view successfully supports it. In addition, further studies by critics Phelan and Sarkissian provide powerful counter-evidence to his view (2009).

In order to test the trade-off hypothesis, Machery presented two pairs of his own vignettes. The first pair are what he calls the 'free-cup case' and the 'extra-dollar case' (2008, p. 179). In both cases, an agent called Joe is feeling dehydrated so he decides to buy the biggest fruit smoothie on offer. In the free-cup case, he is told that if he buys the biggest smoothie he will be given a commemorative cup, while in the extra-dollar case he is told that prices have gone up so he will need to pay an extra dollar for the largest smoothie. In both cases, Joe states that he doesn't care about the extra benefit or the extra cost and buys the smoothie. Participants were asked whether getting the cup or paying the dollar was praiseworthy, blameworthy or neutral and whether they were done intentionally. Participants were found to be significantly more likely to judge that the extra-dollar case was intentional but there was no significant difference regarding the morality question – both actions tended to be judged morally neutral (2008, pp. 180-181).

While Machery argues that this result supports his thesis and challenges Knobe's, the cases are arguably insufficiently similar to the chairman cases to support meaningful comparison. This is primarily because the extra-dollar case arguably does not involve a side-effect at all – paying the extra dollar is not a mere side-effect of buying the smoothie, it is the means by which the smoothie is bought. Since deliberately adopting the means to achieve one's goals is paradigmatically intentional behaviour, the extra-dollar case is not plausibly an intermediate case of intentionality and it does not sufficiently parallel the chairman case. Further, both of Knobe's explanations would predict the results Machery found. For instance, participants are likely to judge that paying an extra dollar is bad while getting a free cup is good, so his original analysis predicts a higher rate of intentionality judgements in the extra dollar case. Since paying an extra dollar is bad and getting a free cup is good, the default positions to which Joe's supposed attitudes are compared are likely to be such that a greater range of intermediate attitudes will be classed as intentional for the extra dollar case, meaning Knobe's later analysis also predicts Machery's results. The smoothie cases thus seem to simply confirm Knobe's analyses.

Machery's second pair of vignettes are equally problematic. Machery attempts to respond to the objection that Knobe's analyses predict the smoothie case results by seeking a result that the trade-off hypothesis, but not Knobe's analysis, would predict. He presented participants with two variants of the trolley problem: the 'worker case', in which an observer John can divert a train to kill one person and save five, and the 'dog case', in which John can divert a train to save five people, kill no one, and incidentally also save a dog. Machery's intention seems to be to provide two cases in which the side-effects are good but only one case (the worker case) involves a trade-off. He asked participants to judge whether either killing the worker or saving the dog were done intentionally. He then asked whether it was 'appropriate' to bring about the side effect in question. Respondents in both conditions typically judged that bringing about the side-effect was appropriate, but that participants were significantly more likely to judge that it was brought about intentionally in the worker case than in the dog case (2008, p. 185).

Machery claims that 'Knobe's account predicts that subjects should judge in *both cases* that the side-effect has not been intentionally brought about', while the trade-off hypothesis only predicts a positive intentionality judgement in the worker case (2008, p. 184, original emphasis). He justifies this claim by pointing out that in both cases, bringing about the side effect was judged to be appropriate. However, as pointed out by Phelan and Sarkissian, Knobe's original hypothesis does not predict a positive intentionality judgement in the worker case since it is highly likely that participants judge killing the worker as bad as well as appropriate, since it is bad to kill him but appropriate to do it to save a greater number of people (2009, pp. 170-171). Further, it can plausibly be assumed that the default attitude towards killing the worker would be something like extreme reluctance, meaning that Knobe's later analysis predicts the result too. If Knobe's hypotheses would have predicted this result too, as seems right, then Machery still lacks evidence that the trade-off hypothesis explains the asymmetry more successfully than Knobe's analysis. Once again, a study designed to undermine Knobe's analyses arguably supports his findings.

Phelan and Sarkissian designed a further study in order to more accurately test the trade-off hypothesis. They presented participants with four adapted versions of a vignette originally presented by Knobe (2003a, p. 192), each describing a lieutenant deciding to send his men into a dangerous situation in which some will be killed in order to maintain control of Thompson Hill. The cases were varied according to two different factors: whether the lieutenant reported caring about the troops or not caring about them, and whether keeping Thompson Hill was described as an important goal in the war or an unimportant one (since the hill was likely to be taken back very quickly). Phelan and Sarkissian argue that if the trade-off hypothesis is correct then the highest rate of intentionality judgements should occur in the Caring/Important case, since this is the clearest example of a trade-off: the lieutenant cares about his men so will see losing them as a cost but he recognises that it is a trade-off he must make to support the war effort (2009, p. 175). However, they found the lowest rate of intentionality judgements in this case, strongly suggesting that the perception of a trade-off is not what causes intentionality judgements in side-effect cases (2009, p. 174).

## *3.4 Values Return: The Deep Self Concordance Model*

The prospect of explaining the asymmetrical intentionality judgements uncovered by Knobe at this point seems fairly bleak. A range of candidate hypotheses have been undermined by experimental results that they cannot explain. This state of affairs has led some critics, such as Phelan and Sarkissian (2009), to conclude that the quest for a parsimonious explanation is a quest pursued in vain. They argue that the results considered above suggest that 'one can arrive at the concept of intentional action by one of any number of disparate routes' so that '[it] is time to abandon the search for parsimony' (2009, p. 179). However, it will here be argued that these philosophers abandoned the search too soon. This is because the statistical analysis of Sripada and Konrath not only challenges Knobe's analyses, as explored above, but also provides a parsimonious way to explain a broad range of the results discussed above. The explanation that they propose draws on the 'Deep Self Concordance Model' first presented by Sripada (2010). The predictions made using the Deep Self Concordance Model are not only borne out by statistical analysis, they also help to make sense of some of the results that baffled previous theories.

### The Deep Self Concordance Model

Chandra Sekhar Sripada suggests that the results found by Knobe's studies can be explained not by moral considerations but by descriptive, non-normative judgements (2010, p. 164). If this is the case – and if his explanation suggests that intuitions about agency are not systematically influenced by some other irrelevant factor – then Sripada's model might yet resolve present worries about using the IAE method to investigate agency. Sripada's theoretical model will here be presented before exploring what experimental data supports it and how well it makes sense of the findings presented above.

At the heart of Sripada's model is the notion that folk psychology includes a 'naïve theory of the structure and contents of the mind' that posits two selves: the 'Acting Self' and the 'Deep Self' (2010, p. 165). The Acting Self is composed of elements typically considered in theories of intentional agency: 'the narrow set of outcome-directed proximal desires, means-end beliefs, and intentions that are the immediate causal source of the action' (2010, p. 165). For instance, when the chairman chooses to adopt the harmful policy his Acting Self is composed of the desire to make profit, the belief that adopting the policy is a way to make profit, and the intention to adopt the policy. The Deep Self, however, is 'a much larger set of more stable, enduring and fundamental attitudes', what Sripada calls 'deep attitudes', that tend to be 'stable and enduring [...] more central to the person's identity and self-conception [...] more abstract [... and] they tend to be reflectively endorsed by the person' (2010, p. 165). Candidate deep attitudes for the chairman might be a stable desire for wealth, an abstract evaluative attitude that a life like that of a successful businessman is the good life, or the belief that pursuing profit is usually more important than protecting the environment.

Sripada describes typical theories of intentional agency as 'Choice/Control Models', according to which making intentionality judgements involves checking acts against an 'inventory' of features relevant to whether the agent chooses her action and whether she controls its occurrence (2010, p. 160). These factors involve the key features of the Acting Self listed above as well as features such as possessing the right skill to control the action. Sripada does not dispute that judgements concerning the features listed in Choice/Control Models are key to intentionality judgements. However, he argues that these judgements are not enough by themselves (2010, p. 164):

> [The] factors that determine whether an agent chooses an outcome and controls its occurrence are not themselves sufficient for concluding whether an [agent] brought about the outcome intentionally. Rather, a crucial additional factor consists of the attitudes contained in the agent's Deep Self, and whether or not those attitudes and the outcome concord.

To put it roughly, Sripada suggests that a key part of the process of making an intentionality judgement is checking whether the action concords with the attitudes of the agent's Deep Self. Sripada describes this as the 'Concordance Criterion', according to which an action is more likely to be judged as intentional when it concords with the attitudes of an agent's Deep Self and it is less likely to be judged as intentional when it is not (2010, pp. 163, 176). It is worth noting that, pace Knobe, this Deep Self Concordance Model insists on the unidirectional thesis. Identifying the Deep Self is, for Sripada, achieved by attributing attitudes to an agent that are stable, enduring, central to the agent's identity and typically reflectively endorsed – none of which need be influenced by normative considerations. Further, checking that an action is controlled, chosen, and concordant with attributed deep attitudes need not involve any normative evaluation of the act. Accordingly, the Deep Self Concordance Model suggests that there is no reason to think that intentionality judgements are influenced by normative considerations.

The Deep Self Concordance Model thus provides a way to dismiss Knobe's claims that intentionality judgements are influenced by moral considerations. In addition, as will be seen below, it has considerable power to explain the results of a range of studies considered here. However, it is worth noting an ambiguity in how the Deep Self Model and the Concordance Criterion are presented. When discussing the model in fairly abstract terms, Sripada sometimes suggests that concordance is a necessary condition for an intentionality judgement to be made. For instance, concordance is described as a 'crucial additional factor' to consider when making an intentionality judgement and the consideration of other factors, such as the agent's desires and skills, is deemed 'insufficient' to inform the judgement (2010, p. 164). When discussing the Concordance Criterion in more detail, however, Sripada claims that consideration of the Concordance Criterion only makes an intentionality judgement more or less *likely*. This suggests that acts may sometimes be judged to be intentional when they are discordant with an agent's deep attitudes, and vice versa. If this is the case, it seems misleading to present concordance as a necessary condition. Quite how we should treat the Concordance Criterion will here be left open, though discussion will proceed on the assumption that only the weaker claim is meant.

Finally, one might be left wondering *why* concordance with deep attitudes would influence intentionality judgements. Sripada's proposal is as follows (2010, p. 166):

> The Deep Self contains the agent's stable, enduring, and most central psychological attitudes, and as a consequence, actions that emerge from an agent's Deep Self are likely to form part of a larger pattern in which actions of this same type regularly and reliably happen again. Actions that are not anchored in an agent's Deep Self are, in contrast, more fleeting and ephemeral, and relatively less likely to form a global, reoccurring pattern. Thus our capacity to predict long-term patterns of behavior requires an ability to distinguish actions that are rooted in the agent's Deep Self from those that aren't. [...] According to the Deep Self Model, the folk concept of intentionality performs precisely this role (that is, in addition to whatever other roles that it performs).

Sripada here suggests that humans have an interest in determining which actions concord with an agent's Deep Self since these are likely to be part of patterns of behaviour over time that, if understood, can help predict and control behaviour. He suggests that the folk concept of intentionality has been shaped partly by this need, so that it tracks which behaviours are concordant with what is taken to be the agent's Deep Self.

If Sripada's Deep Self Concordance Model can satisfactorily explain Knobe's findings then it provides a way to account for them without conceding that intentionality judgements are influenced by moral considerations. This would provide one way to defend the IAE approach adopted in this thesis. However, Sripada's model should not be accepted simply because it supports the approach favoured here. As will be argued below, there are compelling independent reasons to accept the Deep Self Concordance Model. Firstly, the research conducted by Sripada and Konrath provides both convincing evidence for the Deep Self Concordance Model and presents a significant challenge to Knobe's own analysis. Further, the model can explain a wider range of the studies than the previously considered explanations.

## Explanations

The best experimental evidence for the Deep Self Model is that provided by the statistical analysis conducted by Sripada and Konrath (2011). As outlined above, they presented participants with the classic chairman vignettes and asked them six questions designed to establish which of a likely range of factors best explains the asymmetrical intentionality judgements. The order of the questions was systematically varied in order to limit order effects. By using a statistical analysis method called structural path modelling, Sripada and Konrath first established that the normative considerations tested for (the goodness/badness of the outcome and the moral status of the chairman) did not significantly influence intentionality judgements (2011, p. 366). Second, they established that the factors that had the most significant influence on intentionality judgements were 'assessments of the Chairman's underlying values, attitudes and stable behavioural dispositions' – in other words, judgements about his deep attitudes (2011, p. 354). As readers will recall, the fourth survey question asked participants to rank the chairman's values and attitudes towards the environment from 'Very Pro-environment' to 'Very Anti-environment', while the fifth question asked whether the chairman was the kind of person who would bring about similar outcomes in similar situations. Sripada and Konrath argue that, taken together, the answers to these questions reveal attributions of a deep attitude since they test participants' attributions of an attitude that is stable across different situations (2011, p. 359). The answers to these questions had the most significant influence on the intentionality judgements. These two outcomes of their statistical analysis provide substantial evidence in favour of the Deep Self Model – not only do they support the model but they also challenge its rivals.

To explain the results in slightly more intuitive terms, consider how Sripada and Konrath take the Deep Self Model to apply to the chairman case (2011, p. 359):

> [The] model first predicts that in both the harm and help condition, people ascribe to the Chairman core underlying anti-environment values and attitudes. This is because the Chairman says 'I don't care at all about harming [helping] the environment', which is taken to express contempt or hostility towards the environment. In the harm condition, the outcome of harming the environment is concordant with the Chairman's underlying anti-environment

> attitudes, so people say the Chairman brought about the outcome intentionally. In the help condition, the outcome of helping the environment is discordant with the Chairman's underlying anti-environment attitudes so people say the Chairman did not bring about the outcome intentionally.

As can be seen in this passage, the Deep Self Model predicts the asymmetry since if the chairman is judged to have anti-environment deep attitudes in both vignettes then only in the harm condition is the side-effect concordant with his deep attitudes, increasing the likelihood that it is judged as intentional. One might object that the Deep Self Model's analysis rests on the unproven assumption that participants attribute anti-environment attitudes to the chairman in both cases. While attitude attribution results are not explicitly discussed in the structural path modelling study, in a previous study Sripada found that that the chairman was judged to have anti-environment attitudes in both conditions, though he was judged to be significantly more anti-environment in the harm condition (2010, p. 168). This result, alongside the findings of the structural path analysis, makes a compelling case for the Deep Self Model.

Knobe's defender may object that moral considerations might indirectly cause the asymmetry; what attitudes are attributed to the chairman might be influenced by judgements about what attitudes one *should* have towards moral versus immoral outcomes. However, Sripada and Konrath's statistical analysis suggests that this is not the case. Though they did find a small (4%) influence of goodness/badness judgements on judgements of the chairman's attitudes and values, suggesting some normative influence on attitude attributions, they found no significant influence by the participants' own attitudes towards the environment on either the attribution of environmental attitudes to the chairman or on the judgement of how likely he is to produce similar outcomes in similar situations (2011, p. 366). These findings suggest that normative considerations have a negligible indirect influence on intentionality judgements – one too small to explain the asymmetry.

The greatest advantage of the Deep Self Concordance Model is its power to explain a wide range of otherwise baffling cases. For instance, consider two cases with morally neutral outcomes, meaning that any asymmetry is not predicted by the goodness/badness thesis of Knobe's original analysis or any theory that involves blame. First, the Rifle Contest Case is a case from Knobe (2003b) in which an agent really wants to win a shooting contest despite lacking skill with a rifle. He aims for the bull's eye, his hand slips, but luckily he hits it and wins the contest. Participants typically judge that the shooter in this case did not shoot intentionally (2003b).

Second, Sripada's Policeman Rifle Contest Case is the case of an agent who has always wanted to be a policeman but who, when he applies to the local police academy, is asked to compete in an unexpected shooting contest since there are too many applicants. He lacks skill with a rifle and his hand slips, but luckily he hits the target. When Sripada presented this case in a survey, participants typically judged that he did shoot the target intentionally, in contrast to the Rifle Contest Case. Since neither outcome is bad nor blameworthy, and since there is no normative obligation to take any particular attitude towards them, this asymmetry is not predicted by the alternative models considered above. This asymmetry is, however, predicted by the Deep Self Concordance Model since in the Policeman Rifle Contest (but not in the Rifle Contest) participants judge that the outcome concords with the agent's deep attitudes – namely the stable, enduring desire to become a policeman. In contrast, Sripada suggests that in the Rifle Contest Case participants 'infer that the agent must not have any *deep* commitment to winning the contest, since a person who genuinely values winning the contest would have presumably learned how to shoot a rifle' (2010, p. 171). The fact that the Deep Self Concordance model can explain an asymmetry that rival theories cannot counts in its favour.

It might be objected that Sripada's analysis here is a little quick. It does not seem beyond the realm of possibility that participants attribute deep attitudes to the shooter in the Rifle Contest Case, ones that the outcome is concordant with. Participants might judge that the shooter is the kind of person who does rash things to impress others due to a deep desire to show off, one so strong that it overrides other considerations. If participants did attribute a value such and they still judged that his shot was unintentional, this would undermine the Deep Self Model. This objection reveals one of the limitations of the Deep Self Model which is that it may be difficult to have much confidence when identifying the contents of participants' attributions of deep attitudes, if they make any at all. This is because the only attributed attitudes that experimenters can know of are those they ask participants about – a range which may be biased, misleading or limited.

Despite this worry, it seems likely that the Deep Self Model should for now be accepted and the most effective way of testing it developed since, as suggested above, it has greater explanatory power than existing rival accounts. For instance, Knobe's analysis is challenged by his findings concerning the case of Klaus, the selfless Nazi soldier who damages his own army's equipment to save lives – despite the risk of being killed for it. As discussed above, the fact that participants overwhelmingly judge that his act is intentional is difficult for Knobe's analysis to explain since the outcome of Klaus's act is praiseworthy. This case is easily explained by the Deep Self Model, since the fact that he risks his own life to save others plausibly influences participants to attribute a range of deep attitudes concordant with the outcome of saving lives, such as the fact that he "values human life', 'prioritizes the protection of innocents', and 'cares more about saving others than saving himself" (Sripada 2010, p. 171).

The Deep Self Model might also explain the asymmetry in the two bomb vignettes that challenge Knobe's later analysis (2011, pp. 601-602). It seems likely that participants will be more willing to attribute concordant deep attitudes to a man who wishes to detonate a bomb that will kill thousands than to one who wishes to defuse it. Since setting off a deadly bomb is something most people would not wish to do, participants may think it more likely that the desire to do so arises from longstanding attitudes like hatred of a particular kind of lifestyle. On the contrary, most people would like to save thousands of innocents, so this attitude may be taken as less revealing of an agent's deep attitudes which explains the lower rate of positive intentionality judgements in this case.

Similar arguments can be constructed to explain a range of studies presented above. Further empirical study is needed to test them, one that is designed in such a way as to minimise the worry that the attributed attitudes tested for may only represent a biased or narrow range of those that might influence intentionality judgements. However, the statistical analysis performed by Sripada and Konrath, taken together with the significant explanatory power of the model, will here be taken as sufficient reason to accept it.

### Implications

The last question to be addressed is the question of what the implications of the Deep Self Concordance Model are for the present inquiry. First, the model suggests that the IAE approach is safe from the specific threat that Knobe's analysis presents. According to the Deep Self Concordance Model, intuitive intentionality judgements are not influenced by normative considerations but rather by descriptive judgements about features of actions, namely the judgement of whether they concord with the agent's deep attitudes. If intentionality judgements are not influenced by moral considerations, it seems more likely that the IAE approach can be a useful tool for inquiry into what kinds of acts there are.

However, it might be objected that Knobe's findings cannot be dismissed so lightly. It might be argued that concordance with deep attitudes is itself an irrelevant factor that undermines the evidential usefulness of intuitions about agency. If concordance with deep attitudes is only relevant for predicting behavioural patterns, without indicating anything important about what really constitutes intentional agency, then its influence on our intuitions could be just as misleading as the influence of moral considerations. What is really needed to defend the IAE method is not an account of why concordance with deep attitudes is a useful feature for folk psychology, but an account of why it matters for agency itself.

Such an account is arguably what Doris attempts to provide. This can be seen once it is noted how similar Doris's conception of value-expression is to Sripada's notion of Deep Self Concordance. For Doris, an agential, value-expressive act is one caused by a particular kind of desire: those that are longstanding, strong, ultimate and justifying – a desire that structures the agent's behaviour and helps form the narrative of her life. Similarly, Sripada defines Deep Self Concordance as when acts are taken to be in line with a particular kind of attitude: those that are longstanding, central to the agent's identity and reflectively endorsed. Unlike Sripada (who only seeks to describe a folk-psychological notion), Doris attempts to argue that value-expression is what matters for agency and, as was suggested above, his attempt fails. However, the failure of Doris's model need not suggest that all such attempts will not succeed. If an account could be provided that explained how something like Deep Self Concordance helps to constitute the agency distinctive of persons then the IAE method would be placed on firm ground.

Fully articulating such an account is beyond the scope of the present inquiry. However, the work presented here does give some indication of what such an account might look like, as well as what it arguably should not look like. A straightforward response to Sripada's findings would be to argue that it reveals that there are two kinds of human behaviour, the full-blown agency which is instantiated by acts concordant with the attitudes of the Deep Self and other behaviours which are not. This account would face some fairly formidable obstacles, however. For instance, Deep Self theories must make sense of the counter-intuitive way that they imply that two selves are located in one agent. In order to avoid this implication, one might maintain that there is one self with two different kinds of attitudes – deep attitudes and shallow ones. Full-blown agency might then be understood as those acts that concord with an agent's deep attitudes, including concordant acts that are implicitly-governed. However, such a model would be vulnerable to criticisms similar to those made of Doris's account above. Just as Doris seems unable to defend the idea that only value-expressive acts really count as full-blown agency, a Deep Attitude account may struggle to defend the claim that only acts concordant with deep attitudes really count. For instance, consider again the case of agents acting out of character on a fleeting whim, such as the Scrooge who is spontaneously generous to a child. Such acts seem like the acts of the persons who choose to undertake them, even though the attitudes motivating such acts are neither longstanding nor central to the agent's identity. Similarly, akratic acts that are not reflectively endorsed, such as lazily watching television even though one reflectively endorses the idea of exercise, may seem like the act of a conflicted person but this does not defeat the sense that it is an instance of the agency distinctive of persons.

What is needed is an account that can explain what it is about acts that are in line with an agent's values or deep attitudes that makes them instances of full-blown agency without excluding other, everyday instances of apparent full-blown agency from the picture. While a detailed version of such an account must be left for a future project, what is provided here is a gesture to how such an account might begin. It seems plausible that both Doris's account and a Sripada-inspired Deep Attitude account fail because of the way that they explain what is constitutive of full-blown agency by focusing on the involvement of a particular kind of attitude: the kind of desire associated with valuing, or a deep attitude. This strategy fails because so many behaviours that do not involve the special kind of attitude nonetheless strike us as instances of the agency distinctive of persons. Perhaps, instead of attempting to identify a special kind of attitude, an account of full-blown agency should focus on the relation that obtains between the agent and the attitudes that govern her behaviour when an act counts as full-blown agency. Instead of beginning with a consideration of acts and their causes, a fruitful line of inquiry might start by attempting to analyse what kind of relation obtains between the agent and her attitudes when her act expresses her values, when she acts due to autonomous goal control, or when she acts according to reflective deliberation (and so on). Further, such an inquiry might do well by attempting to pin down what different kind of relation obtains when an agent's act is not an instance of full-blown agency, such as when her act is governed by situational control or a compulsive craving.

One possible way to understand the agency-giving relation is suggested by the criticism of Doris above. A notion of self-direction more capacious than mere value-expression seems to be at the heart of the central intuition driving this inquiry, so that an account of self-direction might well be the way to begin. A better understanding of what relation obtains between an agent and her attitudes when an act is self-directed seems key. Perhaps self-direction could be understood as a kind of fitting in with the agent's self-understanding, where this is defined loosely enough to accommodate akratic acts, acting on passing whims and autonomous goal control as well as value-expressive, intentional acts. For instance – to return to the intuitions of Chapter One – a fruitful line of inquiry might begin by closely analysing what relation grounds the sense that an agent's implicitly-governed acts are agential when those acts serve her longstanding goals. Identifying in what exactly the harmony between agent, attitude and act consists in cases like these might both provide the key to a successful account of full-blown agency and give us an account of agency better able to accommodate the influence of implicit cognition.

## Conclusion

This inquiry began by drawing attention to the ways in which some kinds of seemingly agential implicitly-governed behaviour cannot be accommodated by traditional conceptions of agency. Considering these acts uncovered the intuition that what marks out agential implicitly-governed acts is the way in which they align with and arise from an agent's longstanding goals and values. This suggests that one way to account for full-blown agency is to define it as those acts that concord with an agent's values. However, close analysis of John Doris's valuational account of agency revealed the limitations of such an approach.

The failure of Doris's model led to a certain suspicion about the intuition motivating the account. A deflationary response to the intuition would maintain that intuitions about agency ought not to be treated as reliable evidence about it. This led to consideration of the findings of Joshua Knobe. He suggests that our intuitions about agency are influenced by moral considerations, making them unfit to provide reliable evidence about what kinds of agency there are. Close consideration of Knobe's findings revealed that his analysis of them ought to be rejected; what appears at first glance to be the influence of moral considerations turns out to be the influence of ascriptions of deep attitudes to agents. If an agent's deep attitudes can be shown to be relevant to instances of full-blown agency itself, not merely our folk-psychological conception of it, then the method of treating intuitions about agency as evidence can be placed on firmer ground.

Following the above conclusion, it might be thought that a deep attitude concordance model of agency provides the best way to solve the problems considered here. Not only could such an account successfully accommodate instances of autonomous goal control, it could explain and ground our intuition that longstanding values matter for agency. However, such an account would be limited in the same way that Doris's account was found to be. By focusing on the involvement of a particular kind of attitude – a particular kind of desire or a 'deep' attitude – everyday instances of agency that do not involve that attitude are implausibly excluded from full-blown agency. While providing a fully articulated alternative account is beyond the scope of the present inquiry, it has been tentatively suggested that the way to begin setting out such an account would be to attend to the relation that obtains between the agent and her attitudes when an act strikes us as fully agential.

Instead of aiming to develop an account of agency that can accommodate our intuitions about implicitly-governed action, perhaps we would do better to focus on the intuitions themselves. Instead of starting with ideas about agency, we might begin by focusing on what is special about the relation between an agent and her attitudes when implicitly-governed behaviour appears to us to be self-directed. By focusing on the relation of self-direction instead of on the kinds of attitude involved, an account of full-blown agency might be outlined that can better accommodate the pervasive impact of implicit cognition on what we do.

## References

Adams, F. & Steadman, A., 2004. Intentional action in ordinary language: core concept or pragmatic understanding?. *Analysis,* 64(2), pp. 173-181.

Anscombe, G. E. M., 1963. *Intention.* 2nd ed. Oxford: Basil Blackwell.

Bargh, J. A., Chen, M. & Burrows, L., 1996. Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology,* Volume 71, pp. 230-244.

Brownstein, M. & Mavda, A., 2012. Ethical automaticity. *Philosophy of the Social Sciences,* 42(1), pp. 68-98.

Davidson, D., 2001b. Agency. In: *Essays on Actions and Events.* 2nd ed. Oxford: Oxford University Press, pp. 43-61.

Doris, J., 2015. *Talking To Our Selves: Reflection, Ignorance and Agency.* Oxford: Oxford University Press.

Doris, J., Knobe, J. & Woolfolk, R. L., 2007. Variantism about responsibility. *Philosophical Perspectives,* 21(1), pp. 183-214.

Dovidio, J. F. & Gaertner, S. L., 2000. Aversice Racism and Selection Decisions: 1989 & 1999. *Psychological Science,* II(4), pp. 319-323.

Ernest-Jones, M., Nettle, D. & Bateson, M., 2011. Effects of Eyes Images on Everyday Cooperative Behaviour: A Field Experiment. *Evolution and Human Behaviour,* Volume 32, pp. 172-178.

Evans, J. S. B. T., 2008. Dual-Processing Accounts of Reasoning, Judgment and Social Cognition. *Annual Review of Psychology,* 59(1), pp. 255-278.

Evans, J. S. B. T. & Stanovich, K. E., 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science,* 8(3), pp. 223-241.

Fishbach, A. & Shah, J. Y., 2006. Self-Control in Action: Implicit Dispositions Toward Goals and Away From Temptations. *Journal of Personality and Social Psychology,* 90(5), pp. 820-832.

Fowers, B. J., Lyons, E., Montel, K. & Shaked, N., 2001. Positive Illusions about Marriage Among Married and Single Individuals. *Journal of Family Psychology,* Volume 15, pp. 95-109.

Frankfurt, H., 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy,* 68(1), pp. 5-20.

Glaser, J. & Knowles, E., 2008. Implicit Motivation to Control Prejudice. *Journal of Experimental Social Psychology,* Volume XLIV, pp. 164-172.

Gollwitzer, P. M. & Bargh, J. A., 2005. Automaticity in Goal Pursuit. In: A. J. Elliot & C. S. Dweck, eds. *Handbook of Competence and Motivation.* London: The Guildford Press, pp. 624-646.

Haley, K. & Fessler, D., 2005. Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game. *Evolution and Human Behaviour,* Volume 26, p. 245/256.

Hall, L., Johansson, P. & Strandberg, T., 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE,* 7(9).

Hill, G., 1982. Group Versus Individual Performance: Are N+1 Heads Better Than One?. *Psychological Bulletin,* Volume 91, pp. 517-539.

Hyman, J., 2015. *Action, Knowledge & Will.* Oxford: Oxford University Press.

Johansson, P., Hall, L., Sikström, S. & Olsson, A., 2005. Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science,* Volume 310, pp. 116-119.

Knobe, J., 2003. Intentional action and side effects in ordinary language. *Analysis,* 63(3), pp. 190-194.

Knobe, J., 2003. Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology,* 16(2), pp. 309-324.

Knobe, J., 2006. The Concept of Intentional Action: a Case Study in the Uses of Folk Psychology. *Philosophical Studies,* 130(2), pp. 203-231.

Knobe, J. & Mendlow, G. S., 2004. The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology. *Journal of Theoretical and Philosophical Psychology,* 24(2), pp. 252-258.

Korsgaard, C. M., 2008. Acting for a Reason. In: *The Constitution of Agency: Essays on Practical Reason and Moral Psychology.* Oxford: Oxford University Press, pp. 207-230.

Lambert, E. & Dennett, D. C., 2018. Getting by with a little help from our friends. *Behavioural and Brain Sciences,* pp. 29-30.

Levy, N., 2018. Agency is realised by subpersonal mechanisms too. *Behavioural and Brain Sciences,* Volume 41, pp. 30-32.

Liao, S. M., 2008. A defense of intuitions. *Philosophical Studies,* 140(2), pp. 247-262.

Machery, E., 2008. The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind & Language,* 23(2), pp. 165-189.

Malle, B., 2006. Intentionality, Morality, and Their Relationship in Human Judgment. *Journal of Cognition and Culture,* 6(1-2), pp. 87-112.

McConnell, A. R. & Leibold, J., 2001. Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology,* Volume 37, p. 435–442.

Monteith, M., Voils, C. & Nardo, L., 2001. Taking a Look Underground: Detecting, Interpreting and Reacting to Implicit Racial Biases. *Social Cognition,* XIX(4), pp. 395-417.

Moskowitz, G., Gollwitzer, P., Wasel, W. & Schaal, B., 1999. Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals. *Journal of Personality and Social Psychology,* Volume LXXVII, pp. 167-184.

Moskowitz, G. & Li, P., 2011. Egalitarian goals trigger stereotype inhitibion: A proactive form of stereotype control. *Journal of Experimental Psychology,* Volume XLVII, pp. 103-116.

Murray, S., 2018. Why value values?. *Brain and Behavioural Sciences,* Volume 41, pp. 37-39.

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

Nichols, S. & Ulatowski, J., 2007. Intuitions and individual differences: The Knobe effect revisited. *Mind & Language,* 22(4), pp. 346-365.

Payne, K. B., 2006. Weapon Bias: Split-Second Decisions and Unintended Stereotyping. *Current Directions in Psychological Science,* XV(6), pp. 287-291.

Pearson, A. R., Dovidio, J. F. & Gaertner, S. L., 2009. The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology Compass,* 3(3), pp. 314-338.

Pettit, D. & Knobe, J., 2009. The Pervasive Impact of Moral Judgement. *Mind & Language,* 24(5), pp. 586-604.

Phelan, M. & Sarkissian, H., 2009. Is the trade-off hypothesis worth trading off for?. *Mind & Language,* 24(2), pp. 164-180.

Phelan, M. T. & Sarkissian, H., 2008. The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies,* 138(2), pp. 291-298.

Rees, C. F. & Webber, J., 2014. Automaticity in Virtuous Action. In: N. Snow & F. Trivigno, eds. *The Philosophy and Psychology of Character and Happiness.* London: Routledge, pp. 75-90.

Schwartz, D., 1995. The Emergence of Abstract Representations in Dyad Problem Solving. *Journal of the Learning Sciences,* Volume 4, pp. 321-354.

Simonton, D. K., 2004. *Creativity in Science: Change, Logic, Genius, and Zietgeist.* Cambridge: Cambridge University Press.

Snow, N., 2006. Habitual Virtuous Actions and Automaticity. *Ethical Theory and Moral Practice,* 9(5), pp. 545-561.

Sripada, C. S., 2010. The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies,* 151(2), pp. 159-176.

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

Stanovich, K., 2004. Chapter 2: A Brain at War with Itself. In: *The robot's rebellion: Finding meaning in the age of Darwin.* Chicago: University of Chicago Press, pp. 31-80.

Steadman, A. & Adams, F., 2007. Folk concepts, surveys and intentional action. In: C. Lumer & S. Nannini, eds. *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy.* Aldershot: Ashgate Publishing Limited, pp. 17-33.

Ward, Z. B. & Machery, E., 2018. "Defeaters" don't matter. *Behavioural and Brain Sciences,* Volume 41, pp. 52-53.

Webb, T., Sheeran, P. & Pepper, J., 2010. Gaining control over responses to implicit attitude test. *British Journal of Social Psychology 68 ,* pp. 36-51.

Zuckerman, E. W. & Jost, J. T., 2001. What Makes You Think You Are So Popular? Self-Evaluation Maintenance and the Subjective Side of the 'Friendship Paradox'. *Social Psychology Quarterly,* Volume 64, pp. 207-223.