teorema
Vol. XXXV/2, 2016, pp. 7-27
ISSN: 0210-1602
[BIBLID 0210-1602 (2016) 35:2; pp. 7-27]

Situation, Reason and the Extended Agent

Sophie Stammers

RESUMEN

Los hallazgos en ciencia cognitiva muestran que la conducta está regularmente influenciada por rasgos de las situaciones. Se ha mantenido que puesto que los agentes rechazarían esos rasgos en tanto que razones, no actúan libre y agencialmente cuando son influenciados de esta manera. Argumento que este punto de vista está equivocado puesto que (i) descansa en una confusión de tres fundamentos distintos en los que se basa el rechazo de las razones, y (ii) considera una visión a muy corto plazo de lo que constituye un episodio de conducta agencialmente signficativa.

PALABRAS CLAVE: *situacionismo, psicología situacionista, razones, agencia, libre albedrío.*

ABSTRACT

Findings in cognitive science show that behaviour is regularly influenced by situational features. It has been maintained that because agents would reject these features as reasons, they do not act freely and agentially when so influenced. I argue that this view is mistaken, in that it (i) rests on a conflation of three distinct grounds on which to reject reasons; and (ii) takes too short-term a view of what constitutes an episode of agentially significant behaviour.

KEYWORDS: *Situationism; Situationist Psychology; Reasons; Agency; Free Will.*

Findings in cognitive science show that behaviour is regularly influenced by situational features. It has been maintained that because agents would reject these features as reasons, they do not act freely and agentially when so influenced. I show that there are three distinct grounds on which to reject a situational feature as a reason. I then argue that the claim that all situationist experiments show that agents would reject their situational influences is unfounded because it (i) rests on a conflation of these distinct grounds on which to reject reasons; and (ii) takes too short-term a view of what constitutes an episode of agentially significant behaviour. This motivates the need to understand agential responses to situational features in the context of the agent's long-term commitments.

I consider evidence that agents with genuine long-term commitments are able to mediate their responses to at least some situational features, and suggest that this provides a framework for understanding responses to situational features as both reasons-responsive and agential. Consequently, empirical findings may facilitate, rather than limit agency, because they provide an insight into contexts in which agents who do not already have long-term commitments may develop them to engender reasons-congruent responses to situational features.

I. ACTING FOR REASONS

To determine whether any findings in the cognitive sciences threaten free agency, we need to know (i) which conditions must be satisfied in order for free agency to obtain; and (ii) whether the findings in question show that one or more of these conditions is not in fact satisfied. Whilst there is little philosophical consensus on the finer points of the conditions which must be satisfied for free agency to obtain, it has been noted that on many accounts, it is at least a necessary condition that agents are able to act for reasons [Schlosser (2014), pp. 250-251].[1] For Schlosser, the relevant sense of acting for reasons is if the action in question can be *rationalised* from the agent's point of view, which is to say that there is:

> some feature, consequence, or aspect of the action that the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable [Davidson (1963), p. 685, quoted in Schlosser (2014) p. 250].

A number of considerations favour the notion that free agency should at least involve the ability to act for reasons which rationalise action from the agent's point of view. When we consider why free agency is valuable, it is not because we want to be able to perform any action whatsoever — we want to be able to act on the things that *matter* to us, actions for which we see there to be good reasons [Roskies, (2011), Schlosser (2014)]. Further, we tend to care about the kind of agency that grounds moral responsibility, and it has been argued that it is the normative competence of reasons-responsive agents that accounts for why such agents may be held morally responsible for their actions [Wolf (1990), Roskies, (2011)]. A full defence of why the ability to act for reasons has been popularly taken as a necessary condition of free agency is beyond the scope of this paper. Instead, my purpose here is to show that this condition survives a

particular challenge, such as that put forward by Sie and Wouters (2010), and discussed by Nahimas (2006), on the basis of findings in cognitive science.

Before turning to this challenge, it is worth outlining how various philosophers have recruited the notion of acting for reasons to show that a set of findings in neuroscience do not rule out free agency. A series of experiments in a paradigm first designed by Libet and colleagues (1983), and developed by others,[2] apparently show that the conscious intention to make a hand movement is preceded by an unconscious neural signal, known as the 'readiness potential', in the motor cortex of the brain. These results led many neuroscientists to the conclusion that action is determined by unconscious brain activity, that our conscious intentions are not efficacious in bringing about our actions, and that, consequently, we do not have the kind of free agency to which many philosophers are committed.

Experiments in the Libet paradigm continue to be influential in discussions of free agency in the media [Chivers (2010)]. However, a number of philosophers argue that these findings do not threaten free agency precisely because they do not show that our ordinary actions are not done for reasons in the relevant sense [Roskies (2011), Schlosser (2014)]. As Roskies proposes, given the experimental context, there is no reason to prefer to flex at any particular point in the trial over another, and so it is arbitrary *when* a flex occurs [Roskies (2011) p. 18]. In the context of the experiment, flexing now, rather than, say, flexing a moment later, is not an action which could occur for reasons which differ from those that govern flexing a moment later. However, *that* a finger flex occurs at all is not arbitrary, as the agent is responding to her reasons to participate in the experiment and to comply with the experimenter's wishes. So, whilst one might be tempted to argue that lack of conscious initiation defeats the notion that Libet actions are done freely, the fact that Libet actions are arbitrary defeats the notion that Libet actions must turn out to be free for the success of a philosophical account of free agency. Consequently, these findings do not show that we do not regularly exercise free agency.

II. Situational Features

If the ability to act for reasons which rationalise action from the agent's point of view is necessary for free agency, then a multitude of other findings from cognitive science, particularly from cognitive and behavioural psychology, seem apt to pose a challenge to free agency. Sie

and Wouters (2010) make just such a challenge. They summarise a number of experiments from 'situationist psychology' which reveal that aspects of a person's situation, which are not acknowledged as reasons to act by that person, can nevertheless modulate their behaviour, and argue that these findings pose a real problem for accounts of free will committed to acting for reasons [Sie and Wouters (2010)]. Amongst the findings discussed by Sie and Wouters are those which show that subjects who are primed to react with disgust when they encounter an arbitrary word interpret moral transgressions more harshly when the description of the transgression contains one of the primed words [Wheatley and Haidt (2005)]. Additionally, aspects of a person's environment which ought to have no impact on, for instance, the severity of a moral judgement, nevertheless do seem to have an influence. Participants who sit at an unclean table, or in the presence of a bad odour, tend to make harsher moral judgements than those in clean environments [Schnall *et al.* (2008)].

In a further experiment discussed by Sie and Wouters, theology students who were told that they were late as they walked between the two locations of a behavioural study were significantly less likely to assist a stranger slumped on the ground than those who were told that they had a few minutes to spare [Darley and Batson (1973)]. Only 10% of 'late' subjects offered help compared with 63% of 'early' subjects who offered help. This is sometimes called 'the Good Samaritan' experiment. Sie and Wouters note that as the participants were theology students, we would expect them to reject the influence of the situational feature as a reason, arguing that "[s]urely, one would expect people training for a helping profession to be sensitive to the needs of someone in distress, regardless of being in a hurry" [(2010) p.128]. Being in a hurry is not the only situational feature which has been shown to modulate helping behaviour. Further experiments show that whether or not people endeavour to help someone in need is in part a function of the presence and behaviour of other people in the vicinity [Darley and Latané (1968), Latané and Darley (1970)]. These are sometimes called the 'Bystander' experiments. For example, 85% of participants who believed that they were the only person to have heard someone else suffer a seizure over an intercom intervened to help that person, whilst 62% intervened when they believed that there was one other listener, and only 31% intervened when they believed that there were four other listeners [Darley and Latané (1968)].[3] In contrast with the results mentioned in the previous paragraph, participants in these experiments are aware of the situational features, but they generally do not acknowledge the situational features as reasons which

rationalise their behaviour when questioned by experimenters [Darley and Latané (1968), p. 381, Latané and Darley (1970), p. 124].

In addition to the findings acknowledged by Sie and Wouters (2010), a further body of research reveals that many people harbour negative stereotypical attitudes about members of (often marginalised) social groups. For example, Implicit Association Test (IAT) studies, which typically require participants to match pictures of people from different social groups with an evaluative token or a social trait, reveal that participants match stereotype-congruent pairings more quickly than they match stereotype-incongruent pairings [Greenwald *et al.* (1998), Dovidio *et al.* (2007)]. A number of other studies reveal that these stereotypical associations manifest in real-world discriminatory behaviour [see Jost *et al.* (2009) for an overview]. For example, doctors who are shown to harbour negative stereotypical associations on the IAT are less likely to offer treatment to black patients with the clinical presentation of heart disease than to white patients with the same clinical presentation [Green (2007)]. Swedish recruiters who display negative stereotypical associations on the IAT are less likely to offer a job interview to an applicant with a name perceived to be Muslim compared to an equivalent applicant with a Swedish name [Rooth (2007)]. In a video-game simulation, police officers tended to 'shoot' unarmed black suspects at a higher rate than unarmed white suspects [Plant and Peruche (2005)]. Typically, participants do not acknowledge or endorse these negative stereotypical attitudes on questioning, or when explaining their behaviour [Jost *et al.* (2009)].

It has been argued that these results threaten both an account of free will (Sie and Wouters, 2010) and agency (Doris, 2009) on which acting for reasons is a necessary condition. The shared concern is that this wealth of evidence shows that for many actions, if participants had been aware of the situational influences on their actions, then they would have rejected them as reasons which rationalise acting. Sie and Wouters propose that those who think that the ability to act for reasons is a necessary condition for free will must then meet the concern "…that most of our everyday life is determined by automatic processes triggered by external cues…" (2010, p.131) and that "…acting for reasons is exceptional," (2010, p.128). It is further argued that because we are unable to detect whether action is influenced by a factor that we would reject as one of our reasons, we cannot know whether we are acting for reasons or not, and so we also cannot know whether we're acting freely and agentially [Doris (2009) p. 66, Sie and Wouters (2010) p. 128]. I do not intend to respond to this latter claim here, but to the former claim from which it

proceeds — that the data apparently shows that acting for reasons is exceptional. If it turns out that acting for reasons is *not* exceptional, then the latter concern is perhaps less pressing.

III. Distinguishing Grounds on which to Reject Reasons

The foregoing challenge to free agency from situationist psychology might seem problematic for a picture of free agency founded on the ability to act for reasons. But this conclusion would be premature. In this section, I distinguish three grounds on which participants, if aware of the situational influences on their actions, could reject them as reasons which rationalise those actions. I shall then argue that the challenge to free agency rests on a conflation of these distinct grounds on which to reject reasons. Once they are distinguished, it is not clear that all participants really do reject the situational features as reasons in at least some of the experiments, whilst in others, reason rejection may only be a matter of degree.

The first grounds on which to reject a situational feature as a reason is as follows:

> Arational rejection or rejection$_A$: If *S* were to discover the influence of situational feature *f* on her *φ*-ing, she would reject the claim that *f* is the kind of feature that could rationalise *φ*-ing.

This is to say that, regardless of any other reasons *S* has with respect to *φ*-ing, *f* is just not the kind of thing that could count in favour of *φ*-ing. For instance, Alena might reject$_A$ that the colour of her bookshelf is a reason for her to see the film *Avatar*. The colour of her bookshelf is just not the right kind of thing to have any normative or motivational force in favour of watching a particular film, regardless of any other reasons she might have to watch films. We do not need to know about any of Alena's other aims to know that she would reject$_A$ the colour of her bookshelf as a reason here.

The second grounds on which *S* might reject a situational feature as a reason, is if it fails to provide support relative to her other aims. Accordingly:

> Irrelevance rejection or rejection$_I$: If *S* were to discover the influence of situational feature *f* on her *φ*-ing, she would reject the claim that *f* rationalises *φ*-ing *for her*, because it fails to provide

> any normative or motivational force to φ relative to *S*'s other reasons or aims.

For instance, the fact that James Cameron directed the film *Avatar* is not a reason for Kwame to see *Avatar*, because Kwame happens to have no particular aims to see films directed by James Cameron over films directed by anyone else. However, in contrast to rejection$_A$, although Kwame might reject$_I$ some *f* as a reason to φ, it is consistent with his rejection$_I$ that *f* could be a reason for someone *other than* Kwame to φ, where *f* provides normative or motivational force to φ relative to *their* aims. For instance, for Chloe, a loyal James Cameron fan who aims to see all films directed by James Cameron, the fact that James Cameron directed the film *Avatar* is a reason for Chloe to see it.

Thirdly, *S* might accept that some situational feature rationalises an action to an extent, but reject that it is a sufficient reason for that action, because it is defeated by other reasons not to perform the action. As such:

> DEFEATER REJECTION OR REJECTION$_D$: If *S* were to discover the influence of situational feature *f* on her φ-ing, she would reject that *f* *sufficiently* rationalises φ-ing *for her*, because whilst it might provide some normative or motivational force to φ, this is defeated by *S's* other reasons *not* to φ.

For instance, Sunil might accept that impressive special effects are a reason to see *Avatar,* but reject that they are a sufficient reason, in light of his aim to only watch films which pass the Bechdel test,[4] which he prioritises over seeing films with impressive special effects. We might think of rejection$_I$ and rejection$_D$ as alike, in that both are indexed to the agent's other aims and reasons, and that we cannot simply assume that agents will reject$_I$ or reject$_D$ any feature without knowing about their other aims.

Would participants reject$_A$, reject$_I$ or reject$_D$ the situational features as reasons, which are nevertheless shown to influence them in the foregoing experiments? The answer differs, depending on which experimental paradigm is in question, as I will argue in the following.

Arational rejection

In at least some experiments, it seems that rejection$_A$ characterises the way in which agents would reject the situational influences: Schnall and coauthors suggest that participants would not endorse situational features such as a nearby dirty table or bad smell as a reason to judge

moral vignettes more harshly, as they were "'tricked' by *extraneous* disgust" into making harsher judgements [(2008) p. 1107, emphasis mine]. Indeed, the hygienic conditions of the surroundings in which one comes to a moral judgement about a hypothetical scenario are extraneous, and bear no normative relation to the moral merits of the case being judged, regardless of the aims of the agent. Similarly, Wheatley and Haidt suggest that prime words are 'arbitrary' modulators of moral judgements [(2005) p. 783]. An induced association between a particular word and a feeling of disgust bears no normative relation to the moral features of the vignette at issue, regardless of the aims of the agent. If we assume that people would only endorse rationalising features which are able to provide normative or motivational force to the judgment in question, then we're entitled to interpret these results as situations in which the participants would reject$_A$ that the situational features shown to influence them rationalise their judgements. Accordingly, participants will not endorse arbitrarily primed words or smells as reasons for them to make harsher moral judgements.[5]

There are a number of ways that philosophers have interpreted experiments which apparently show that participants would reject$_A$ situational influences as reasons. For instance, Sandis (2015) proposes that situational features do not replace or invalidate the reasons for which participants take themselves to be acting. Instead, what situational features do is to increase the salience of participants' reasons. He suggests that if participants are asked not what their reasons for acting are, but *why* those reasons are salient to them, then they might well give different answers, implying that they might possibly name situational features [(2015) p.270]. So, participants would not reject$_A$ the reasons for which they acted, because these considerations really do rationalise their actions.

I agree with Sandis that situationist findings do not show that our reason talk is fundamentally confused, but I wonder whether there is still a problem for agency that arises at the level of reasons-salience, rather than at the level of reasons-responsiveness: We want our reasons to be salient because they track facts, not because they are made salient by situational features. That is to say that if, for instance, a jury concludes that Defendant One's actions are more serious than Defendant Two's actions, then we would want the legal significance of the first defendant's actions to be salient to jurors because they *were* more serious than Defendant Two's actions, not because there was a bad smell in the room at the time of hearing Defendant One's account.

Another way to interpret the situationist challenge, particularly when it comes to arational rejection is that an agent's capacity for acting for reasons comes in varying degrees. Nahimas (2006) has argued that agency, construed as reasons-responsiveness, is a property that agents possess and exercise in varying degrees. He points out that this is somewhat intuitive and fits with our notion of young children as developing agents, who do not yet meet the necessary conditions to be considered fully morally responsible for all of their actions [Nahimas (2006) pp. 171-2]. Similarly, adults may be responsive to reasons in varying degrees. For instance, the harshness of a person's judgement of a moral vignette may be the combined result of (i) the moral reasons that favour judging the characters therein harshly, and (ii) the situational features which the agent would reject$_A$ as reasons. To the extent that they respond to the moral reasons evoked in the vignette, they judge freely and agentially, and to the extent that their judgement is modulated by situational features which do not rationalise the judgement, they judge non-agentially. Just because an agent would reject$_A$ some situational feature that modulates a judgement as a reason, this does not mean that they judge wholly non-agentially, but rather that they judge partially non-agentially. Recall that in the experiments of both Wheatley and Haidt (2005), and Schnall *et al.* (2008), agents do not act wholly on the basis of the situational features, but also on the basis of the moral features of the vignettes. So, this kind of evidence does not warrant the conclusion that we rarely ever act for reasons, just that our capacity to do so might sometimes be a matter of degree. Developments in cognitive science, then, enable us to enhance our agency by manipulating decision environments in order to shield judgements from the influence of unendorsed situational features. For instance, a parent scolding a child might manoeuvre the argument so that it takes place away from the currently dirty kitchen, so that their moral judgements are not unduly severe.

Irrelevance and defeater rejection

When it comes to the Bystander, Good Samaritan and implicit bias results, rejection$_A$ does not seem to characterise the way in which participants would reject their situational influences as reasons. Here, the situational features which were shown to influence action *could* count as reasons for acting, depending on the participant's aims. Being in a hurry is the right kind of thing to rationalise walking past someone in need for a person who values punctuality at events to which they have already committed over benevolence to a stranger. Knowing that four other

people witnessed a person having a seizure in another room is the right kind of thing to rationalise inaction for someone who is more concerned with not making a fool of themselves in a situation of uncertainty, than with offering help. Knowing that a job candidate is black is the right kind of thing to rationalise not inviting that person to interview for an explicitly racist recruiter. So, if participants in these experiments reject the situational features as reasons, then it is because they either do not possess the aims which make the situational features relevant (in which case they reject$_I$ the situational features as reasons to act) or they possess reasons which trump the normative force of the situational features (in which case they reject$_D$ the situational features as (sufficient) reasons to act). Since it is possible that the situational features *could* rationalise action in these experiments, as compared with the priming and environmental hygiene experiments, we can only conclude that participants would reject such features as reasons if we know that participants do not possess any aims in light of which the situational features rationalise acting. In what follows, I argue that it is not clear that we can glean this information from the experimental results.

One might think that we know that the situational features do not rationalise participants' actions from their point of view because participants do not acknowledge these features as reasons when experimenters ask what influenced their behaviour [according to Darley and Latané (1968), p. 381, Latané and Darley (1970), p. 124]. And since participants do not acknowledge the situational features as reasons, it might be concluded that participants would either reject$_I$ that such features have any relevance for them (even though they might be relevant to someone else) or reject$_D$ that such features provide sufficient reason to fail to help, for instance. But there are a number of problems with this line of argument. Firstly, subjects provide an account of what influenced them *after* the experimental trial, rather than at the time of action. It is therefore wrong to claim that participants' reports of what they took to influence them rule out that they took the situational features as reasons which counted in favour of action (or inaction) at the time of acting. At most, these reports show that participants do not have *retrospective* awareness that they took the situational features as reasons for action [Nahimas (2006) p.179]. Participants report the memory of what they believe influenced them, rather than reporting their experiences of the influences as they occurred. However, the memory of the influence of a situational feature, especially one that paints the participant in a morally negative light, may

well have been lost or overwritten by the time participants report what they took to influence them.

Relying on participants' post-trial reports is problematic for another reason. There is evidence to suggest that people alter their reports of their personal attitudes depending on the conditions in which they are asked to report these attitudes. For instance, it has been shown that participants express more highly prejudiced racial attitudes on a self-report questionnaire when their responses are anonymous as compared to when their responses are reported to the experimenter [Plant and Devine (1998)]. One way to interpret these results is that participants are more highly motivated to be seen to believe and behave in line with the socially desirable norms of non-prejudice when they know that they have to hand in their attitude report in person, compared to when their reports are anonymous. Other results show that lack of motivation to comply with socially desirable norms modulates self-reports of prejudice. Generally, participants tend to report attitudes on prejudice questionnaires that do not correlate with their Implicit Association Test (IAT) scores — usually implicit attitudes are more prejudiced than self-reports [Nosek *et al.* (2007)]. However, if participants are led to believe that experimenters will *know* if their self-reports do not match their 'genuine' personal attitudes, (thus reducing the motivation to report socially desirable attitudes in place of accurate personal attitudes) then they report personal attitudes which *do* correlate with their IAT scores [Nier, (2005)]. This suggests that participants' own reports of what they take to influence them cannot always be taken at face value, particularly if participants are motivated to conceal that they may be moved by particular considerations, in order to give a more socially desirable response. It is possible that at least some participants in the Good Samaritan or Bystander experiments do recognise the situational features as sufficient reasons to hurry to their talk,[6] or to refrain from possibly making a fool of themselves, but that when it comes to reporting these to experimenters, the pressure to comply with moral norms, and not to acknowledge oneself as an ostensibly bad person, mediates their responses.[7]

This is not to say that all participants in the foregoing experiments *did* recognise and act on the situational features as reasons, but that because these situational features *could* rationalise their actions, and because of the foregoing problems in determining the accuracy of a retrospective report, we cannot say incontrovertibly that they didn't. Still, one might think that this doesn't really vindicate free agency because whilst we end up with a picture on which we act for reasons, we are much more forget-

ful, self-deceived, self-serving and prejudiced than we would perhaps ideally like to be. One might further think that the picture of agency we're left with is not particularly desirable if it does not ground moral responsibility. In the remainder of the paper, I suggest that this conclusion is premature, and is based on taking too narrow a view of what constitutes a particular episode of agentially significant action. I argue that if we see agency in the context of long-term commitments, then we may understand responses to situational features as agential after all.

IV. Responding to Reasons in an Extended Temporal Context

Consider that many of the situationist experiments investigate agency in a fairly discrete and limited temporal stretch. This is true of both the Good Samaritan and the Bystander experiments, both of which investigate one instance of helping behaviour. The influence of the situational feature (together with participants' failure to report their endorsement of this feature as rationalising) is taken as chiefly explanatory of why participants do not provide assistance to a person in need. However, even if people fail to act on reasons to help someone in need, this doesn't necessarily mean that being the kind of agent who fails to provide assistance in a particular circumstance isn't itself the result of an agential process.

In fact, it is illuminating to consider that the episodes of behaviour that are most instructive in these experiments are in fact *omissions*, where participants *fail* to provide the help which we think is morally required. There is a sizeable philosophical literature on omissions, and on whether or not omissions are culpable [for example, see H. Smith, (1983), A. Smith (2005), Sher (2006)]. Whilst there is disagreement as to exactly which conditions ground culpability, all accounts agree that simply looking at the omission itself, in isolation of any other factors, is insufficient to determine whether or not the omission is agential and the agent is culpable.

Consider the following example. Siblings Emma and Sarah receive an email from their third sibling, Julie, who is caring for their elderly father. The email reveals that their father will undergo a serious operation on the 23rd June, and that they ought to do their best to clear their schedule so that they are able to visit him in hospital on that day. Emma cares about her father a lot, and on hearing the news writes the date in

her diary, sets a reminder in her email calendar, and makes a note to book time off work. When Emma glances at her diary and sees the note of the upcoming operation she thinks about buying flowers, and then remembers, after visiting her partner's father, that hospitals don't actually allow flowers inside wards anymore. Instead, she thinks about making sure that she brings some newspapers for her father to read as he recovers. Sarah, who does not much care for her father, neglects to make a note of the date of the operation and doesn't think about it again. June 23rd arrives, and Emma meets Julie in the hospital. Sarah doesn't turn up — she has forgotten the date.

If we look just at events on June 23rd, then it is difficult to fully appreciate the character of Sarah's failure to turn up at the hospital. But in the context of the antecedent conditions which preceded this omission – a failure to even make a note of the date – the omission starts to look more like a failure of care and commitment than a simple lapse of memory. It is not clear that just pointing to the fact that on June 23rd, Sarah forgot to visit her father fully characterises the omission. There is a further reason that she failed to visit which explains why she was unresponsive to reasons to visit her father on June 23rd: she simply didn't care enough to make the basic preparations necessary for remembering the date. So, to determine whether or not the omission is a failure of care or just a lapse of memory, we need to know something about the agent's attitudes in the antecedent circumstances which led to the omission.

Neither the Good Samaritan nor the Bystander experiments give us this kind of context. But, arguably, this kind of context matters. Unless we are able to guarantee that participants in fact *do* reject_{I} or reject_{D} the influences of the situational features (which I raised doubts about in §III), the conclusion that people tend to be more self-serving than we normally suppose is as consistent with the data as the conclusion that people lack any sort of self-serving goals, but respond unwittingly to situational features as if they did. Whilst it is true that much variation in the situationist studies is explained by the influence of the situational features, there is still variation between the helpers and the non-helpers in each condition that remains unexplained, possibly leaving room for personal commitments to play some role.

To this suggestion, it might be objected that in both the Bystander and Good Samaritan studies there were no observed correlations between people who reported to be helpers on personality questionnaires and people who helped the person in need. For example, in the Bystander experiments, participants responded to measures which indicate the

extent to which they agree with statements such as 'I am the kind of person people can count on' and 'I would never let a friend down when he expects something of me' [from Berkowitz and Daniels' 'Social Responsibility Scale' (1964)]. There are two issues here. Firstly, we might think that the relevant personal commitments that might explain behaviour are not whether participants take themselves to be *generally* helpful people, but whether they value helping a stranger over helping a person to which they have already committed (by being on time for a presentation, for instance) and it is not clear that the cited measures give this fine a grain of detail. Secondly, self-report measures of personality traits are as open to modulation by social desirability concerns as self-report measures about reasons for action. A competing hypothesis might explain the lack of correlation: a general trait of self-interest could potentially produce both a failure to help and reports that one is a nice person on a personality measure.

Again, this is not to say that personality questionnaires never reveal long-standing commitments, or never correlate with commitment-congruent behaviour in the context of situational influences. Indeed, in the case of implicit bias, it has been shown that professing to hold a long-standing commitment on a self-report questionnaire *does* correlate with reduced manifestation of implicit prejudice. It turns out that agents who report that they care about the implications of prejudice, and endorse non-prejudice because they think it is an inherently good thing, manifest less implicit bias than those who report that they endorse non-prejudice for approval from others, or those who report that they do not have any long-standing commitments to egalitarian behaviour [Monteith, Sherman and Devine (1998), Devine *et al.* (2002)].

Nevertheless, the lack of a correlation between personality questionnaires and behaviour in the context of situational influences does not *guarantee* that helping or omitting to help is not the result of something agential. There are other measures besides self-report questionnaires which determine agential commitments, and which are not so open to mediation by social desirability concerns. It has been shown that when people are made to engage in behaviour which violates what they report to be a genuinely held long-term commitment, they try to alleviate the conflict felt by overcompensating later on, performing what have been termed 'incompleteness behaviours' — behaviours which reflect the commitment in question [Gollwitzer, Wicklund and Hilton, (1982)]. As measuring long-term commitments in this way requires that the commitments in question are *exercised* to bring about commitment-congruent behaviour, it provides

a more direct record of the commitment at issue than measures which rely solely on participants' own reports about their commitments [Moskowitz *et al.*(1999) p. 169]. Speaking of long-term commitments to non-prejudice, Moskowitz *et al.* maintain that:

> If people commit themselves to such self-defining goals, they are expected to make use of available opportunities to express the goal and to hold on to it even in the face of hindrances, barriers, and difficulties [Moskowitz *et al.* (1999) p. 169].

Accordingly, Moskowitz *et al.* (1999) wanted to know whether agents who perform egalitarian incompleteness behaviours after participating in a non-egalitarian task would also tend to manifest less implicit bias. If this was the case, then perhaps long-term egalitarian commitments enable agents to produce implicit behaviours in line with their egalitarian goals. As hypothesised, they observed a correlation between those who performed incompleteness behaviours after engaging in a non-egalitarian task (so, those with long-term commitments to egalitarianism) and those who manifested less implicit bias [Moskowitz *et al.* (1999)].

Interestingly, Moskowitz and co-authors maintain that the processes which bring implicit responses in line with an agent's long-term commitments are not conscious, or effortful, and do not require the agent to consider whether they still believe that they should refrain from prejudice whenever the relevant social concepts are made salient. Moskowitz *et al.* propose that the long-term egalitarian commitment in question operates automatically to prevent the facilitation of stereotypic categories in the presence of the relevant social concepts [(1999) p. 168]. Accordingly, agents who have cultivated longstanding commitments to egalitarianism have done so as a result of *already* having responded to reasons to refrain from prejudice. Such agents, it turns out, do not need to consciously consult these reasons each time they find themselves in a situation where they could act prejudicially or fairly, in order for the commitment to engender reasons-congruent behaviour.

One might think that if an agent's longstanding commitment is non-effortfully activated to produce egalitarian actions, then these actions are not done for reasons, and are not agential. But this seems unfounded. Consider Steup in the following:

> I'd like to see the person who, just before brushing her teeth, forms the intention to unscrew the cap of the toothpaste tube. But surely unscrewing the cap of one's toothpaste is not an unfree action [Steup (2008) p. 383].

Steup formulates the issue with respect to intentions, but I think that the point is equally applicable to reasons-responsiveness. Just because a person does not consciously recognise any reasons to unscrew the toothpaste cap, it does not mean that she does not unscrew the cap for reasons—reasons that she would endorse as rationalising her action if she were asked. Moskowitz *et al.*'s (1999) findings show that whilst individuals may not recognise that situational features such as (their perception of) a job interviewee's race affect their judgements *as* they make them, whether or not they are the kind of person to be influenced by such situational features in the first place is at least in part determined by their long-term commitments — attitudes which are both reasons responsive and agential.

Holroyd and Kelly (2016) maintain that the utilisation of long-term commitments to calibrate responses to situational features in line with the agent's values is rightly considered as an agential capacity, even when such responses are not guided by attention, as is typical of implicit bias. Accordingly:

> The agent's values and goals themselves, then, can play a role qua mechanisms that influence and calibrate the subsystems that run without reflective or direct control. This is a case of one element of a person's psychological economy influencing another. The agent's values 'keep in check' the operation of implicit bias, such that pursuing certain values is one way of exercising ecological control even when one is not actively monitoring one's actions with respect to whether they promote (or depart from) those values. Crucially, this can be so without the agent expressly intending, at any point, to put in place mechanisms for this purpose [Holroyd and Kelly (2016) p. 123].

Holroyd and Kelly formulate the issue with respect to actions in line with an agent's *values*, but we might think that an agent develops their values because of the reasons they see to endorse such values.

It might be thought that having to cultivate long-term commitments in order to act for reasons that one endorses results in a rather demanding account of free agency. However, Holroyd and Kelly suggest that the cultivation of a capacity to respond agentially without the need for deliberation – as well as a number of other cultivated control strate-

gies that I do not have space to outline here – are actually frequently employed in everyday action guidance. They illustrate this with the example of a tennis player who responds instinctively, without deliberation, to a baseline shot. Simply because the tennis player did not have to deliberate in the moments before taking the shot in order play effectively, the shot is no less a product of their agency [Holroyd and Kelly (2016) p. 119]. Consequently, Holroyd and Kelly maintain that this type of control is in fact rather mundane, and "underlies a vast swathe of human behaviour and problem-solving" [(2016) p. 123].

So, to fully appreciate reasons-congruent responses to situational features, as well as failures to screen their influence, the relevant time-frame of enquiry must encompass all of the attitudes, long-standing commitments included, which constitute a particular episode of agentially significant action. It is not clear that the Good Samaritan or Bystander experiments give us an account of these attitudes, with sufficiently reliable measures. But we can observe how long-term commitments, once cultivated, may mediate reasons-congruent responses to situational features, without the need to consciously consult one's reasons, with regard to implicit bias. Therefore, it would be premature to conclude that because of findings in situationist psychology, we do not act freely and agentially when influenced by situational features that we wouldn't endorse as reasons. Acting freely and agentially, whether it is unscrewing the toothpaste cap whilst deep in thought about something else, performing implicitly egalitarian behaviour, or responding to a person in need, may well be a matter of cultivating the relevant long-term commitments.

CONCLUSION

Findings in cognitive science may show that behaviour is regularly influenced by situational features that, some have claimed, agents would reject as reasons. However, whether participants really would reject these features as reasons depends on whether the features in question can or cannot rationalise action, and on whether they are irrelevant to (or defeated by) participants' other reasons and commitments. I argued that a number of these results do not incontrovertibly demonstrate that participants reject situational features as reason-giving, because it was not shown that participants lack other relevant self-serving goals or commitments, whilst other results only show that reasons-responsiveness is diminished, rather than eliminated. Agentially significant behaviour can

extend over time, and agents with genuine long-term commitments are able to mediate their responses to at least some situational features. I made the case that responses mediated by long-term commitments are correctly modelled as reasons-responsive and agential. Viewing agents' rational capacities as extended in this manner both vindicates agency, and illuminates the role that cognitive science can play in fine tuning, rather than restricting agency.

Department of Philosophy
King's College London
Strand, London WC2R 2IS, UK
E-mail: sophie.stammers@kcl.ac.uk

NOTES

[1] A significant number of both compatibilists and incompatibilists about free will accept this condition [for discussion, see Schlosser (2014) p. 251].

[2] For instance, see Haggard et al. (2002), Soon et al. (2008).

[3] For brevity, when I mention the 'Bystander experiment' in the singular, it is to this particular experiment that I mean to refer.

[4] To pass the Bechdel Test, a film must feature two female characters who have a conversation with each other about something other than a man. Many classic blockbuster films fail the test.

[5] One might think that it is not impossible that an agent might cultivate an aim to make harsher moral judgements when in the presence of a bad odour, and if so, then they might endorse a bad odour as a reason to judge harshly. But it seems unlikely that a sufficient number of participants would just happen to have this specific aim such that the results are invalidated.

[6] Indeed, the Good Samaritan experimenters consider this very possibility on the final page of the paper as a possible alternative hypothesis, but this appears to have been overlooked by many who discuss their results [Darley and Batson (1973) p. 108].

[7] In fact, the Bystander experimenters consider both the possibility that subjects might misremember their reasons, and the possibility that they might misreport socially undesirable rationalisations, but dismiss them, maintaining "it is our impression, however, that most subjects checked few [options of what crossed their mind during the apparent seizure] because they had few coherent thoughts during the fit" [Darley and Latané (1968), p. 381]. It is not clear why being flustered or confused during witnessing the seizure is incompatible with taking the number of people on the call as a reason not to help, and then failing

to report this — potentially, if one is both confused, and knows that there are others on the line, then this might further rationalise not helping.

REFERENCES

BERKOWITZ, L. and DANIELS, L. (1964), 'Affecting the Salience of the Social Responsibility Norm: Effects of Past Help on the Response to Dependency Relationships', *The Journal of Abnormal and Social Psychology,* vol. 68(3), pp. 275-281.

CHIVERS, T. (2010), 'Neuroscience, Free Agency and Determinism', *The Telegraph,* 12th October, URL*: http://www.telegraph.co.uk/news/science /8058541/ Neuroscience-free-will-and-determinism-Im-just-a-machine. html,* accessed on 02/03/16.

DARLEY, J. M. and BASTON, C. D. (1973), '"From Jerusalem to Jericho": A Study of Situational and Dispositional Variables in Helping Behavior', *Journal of Personality and Social Psychology*, vol. 27(1), pp. 100-108.

DARLEY, J. M. & LATANÉ, B. (1968), 'Bystander Intervention in Emergencies: Diffusion of Responsibility', *Journal of Personality and Social Psychology*, vol. 8(4), pp. 377-383.

DASGUPTA, N. and GREENWALD, A. G. (2001), 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals', vol. 81(5), pp. 800-14.

DAVIDSON, D. (1963), 'Actions, Reasons, and Causes', *Journal of Philosophy*, vol. 60, pp. 685-700.

DEVINE, P. G., PLANT, E. A., AMODIO, D. M. *et al.* (2002), 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice', *Journal of Personality and Social Psychology*, vol. 82(5), pp. 835-848.

DORIS, J. M. (2009), 'Skepticism About Persons', *Philosophical Issues,* vol. 19(1), pp. 57-91.

DOVIDIO, J. F., KAWAKAMI, K., JOHNSON, C. *et al.* (2007), 'On the Nature of Prejudice: Automatic and Controlled Processes', *Journal of Experimental Social Psychology,* vol. 33(5), pp. 510-540.

GOLLWITZER, P. M., WICKLUND, R. A. and HILTON, J. L. (1982), 'Admission of Failure and Symbolic Self-Sompletion: Extending Lewinian Theory', *Journal of Personality and Social Psychology*, vol. 43(2), pp. 358-371.

GREEN, A. R., CARNEY, D. R., PALLIN, D. J. *et al.* (2007), 'Implicit Bias Among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients', *Journal of General Internal Medicine,* vol. 22(9), 1231-1238.

GREENWALD, A. G., MCGHEE, D. E. and SCHWARTZ J. L. K. (1998), 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test', *Journal of Personality and Social Psychology*, vol. 74(6), pp. 1464-1480.

HAGGARD, P., CLARK, S. and KALOGERAS, J. (2002), 'Voluntary Action and Conscious Awareness', *Nature Neuroscience,* vol. 5(4), pp. 382-385.

HOLROYD, J. and KELLY, D. (2016), 'Implicit Bias, Character, and Control', in. Webber, J. and Masala, A. (eds.), *From Personality to Virtue*, Oxford, Oxford University Press, pp. 106-133.

JOST, J. T., RUDMAN, L. A., BLAIR, I. V. *et al.* (2009), 'The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that No Manager Should Ignore', *Research in Organizational Behavior,* vol. 29, pp. 39-69.

LATANÉ, B. and DARLEY, J. M. (1970), *The Unresponsive Bystander: Why Doesn't He Help?* Englewood Cliffs, N.J., Prentice-Hall.

LIBET, B., GLEASON, C. A., WRIGHT, E. W., *et al.* (1983), 'Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (readiness-potential). The Unconscious Initiation of a Freely Voluntary Act', *Brain*, vol. 106(3), pp. 623-642.

MONTEITH, M. J., SHERMAN, J. W. and DEVINE, P. G., (1998), 'Suppression as a Stereotype Control Strategy', *Personality and Social Psychology Review*, 2(1), pp. 63-82.

MOSKOWITZ, G. B., GOLLWITZER, P. M. and WASEL, W. *et al.* (1999), 'Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals', *Journal of Personality and Social Psychology*, vol. 77(1), pp. 167-184.

MOSKOWITZ, G. B. and LI, P. (2011), 'Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control', *Journal of Experimental Social Psychology,* vol. 47(1), pp.103-116.

NAHIMAS, E. (2006), 'Autonomous Agency and Social Psychology', in Marraffa, M., De Caro, M. and Ferretti, F. (eds.), *Cartographies of the Mind,* Dordrecht, Springer, pp. 169-185.

NIER, J. A. (2005), 'How Dissociated Are Implicit and Explicit Racial Attitudes? A Bogus Pipeline Approach', *Group Processes & Intergroup Relations*, vol. 8(1), pp. 39-52.

NOSEK, B. A., GREENWALD, A. G. and BANAJI, M. R. (2007), 'The Implicit Association Test at Age 7: A Methodological and Conceptual Review'; in Bargh, J. A. (ed.), *Automatic Processes in Social Thinking and Behaviour*, New York, Psychology Press, pp. 265-292.

PLANT, E. A. and DEVINE, P. G. (1998), 'Internal and External Motivation to Respond Without Prejudice', *Journal of Personality and Social Psychology*, vol. 75(3), pp. 811-832.

PLANT, E. A. and PERUCHE, M. (2005), 'The Consequences of Race for Police Officers' Responses to Criminal Suspects', *Psychological Science,* vol. 16(3), 180-183.

ROOTH, D. (2007), 'Implicit Discrimination in Hiring: Real World Evidence', *IZA Discussion Paper,* No. 2764, Bonn, Forschungsinstitut zur Zukunft der Arbeit.

ROSKIES, A. (2011), 'Why Libet's Studies Don't Pose a Threat to Free Will', in Sinnott-Armstrong, W. and Nadel, L. (eds.), *Conscious Will and Responsibility,* New York, Oxford University Press, pp. 11-22.

SANDIS, C. (2015), 'Verbal Reports and "Real" Reasons: Confabulation and Conflation', *Ethical Theory and Moral Practice,* vol. 18(2), pp. 267-280.

SCHLOSSER, M. E. (2014), 'The neuroscientific study of free will: A diagnosis of the controversy', *Synthese*, vol. 191(2), pp. 245-262.

SCHNALL, S., HAIDT, J., CLORE, G. L. *et al.* (2008), 'Disgust as Embodied Moral Judgment', *Personality and Social Psychology Bulletin,* vol. 34(8), pp. 1096-1109.

SHER, G. (2006), 'Out of Control', *Ethics,* vol. 116(2), pp. 285-301.

SIE, M. and WOUTERS, A. (2010), 'The BCN Challenge to Compatibilist Free Will and Personal Responsibility', *Neuroethics,* vol. 3(2), pp. 121-133.

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

SMITH, H. (1983), 'Culpable Ignorance', *Philosophical Review,* vol. 92(4), pp. 543-571 .

SOON, C. S., BRASS, M., HEINZE, H.-J. *et al.* (2008), 'Unconscious Determinants of Free Decisions in the Human Brain', *Nature Neuroscience,* vol. 11, pp. 543-545.

STEUP, M. (2008), 'Doxastic Freedom', *Synthese*, vol. 161(3), pp. 375-392.

WHEATLEY, T. and HAIDT, J. (2005), 'Hypnotic Disgust Makes Moral Judgments More Severe', *Psychological Science,* vol. 16(10), pp. 780-784.

WEBB, T. L., SHEERAN, P. and PEPPER, J. (2012), 'Gaining Control Over Responses to Implicit Attitude Tests: Implementation Intentions Engender Fast Responses on Attitude-Incongruent Trials', *British Journal of Social Psychology*, vol. 51(1), pp. 13-32.

WOLF, S. (1990), *Freedom Within Reason,* New York, Oxford University Press.