dialectica Vol. 71, N° 2 (2017), pp. 209–229
DOI: 10.1111/1746-8361.12177

Reasons Internalism and the Function of Normative Reasons

Neil Sinclair[†]

Abstract

What is the connection between reasons and motives? According to Reasons Internalism, there is a non-trivial conceptual connection between normative reasons and the possibility of rationally accessing relevant motivation. Reasons Internalism is attractive insofar as it captures the thought that reasons are for reasoning with and repulsive insofar as it fails to generate sufficient critical distance between reasons and motives. Rather than directly adjudicate this dispute, I extract from it two generally accepted desiderata on theories of normative reasons and argue that a new theory can satisfy both. The new theory locates part of the meaning of normative reason statements in their role in normative discussion. It generates a view of the connection between reasons and motives that is distinct from Reasons Internalism, yet distinctively in its spirit.

1. *Reasons Internalism*

What is the connection between reasons and motives? More precisely, what connection is there, if any, between it being true that A has a normative reason to Φ and the possibility of A being motivated to Φ? According to *Reasons Internalism* there is a non-trivial conceptual connection of the following form:

> RI: Necessarily, for all agents A, circumstances C, and actions Φ, A in C has a normative reason to Φ only if A in C can rationally access the state of being motivated to Φ from existing motivations.

RI is one way of capturing the thought that agents' reasons are constrained by their actual motivations. The constraint is indirect: agents' reasons are directly constrained by the outcomes of processes of rational deliberation, but those outcomes in turn are constrained by agents' actual motivations. This is in contrast to the so-called sub-Humean model, which takes reasons to be directly constrained by agents' actual motivations (Williams 1981, 101–2; Schroeder 2007). In some form, RI is accepted by Brandt (1979), Williams (1981, 1989, 1995), Smith (1995), Setiya (2007, 2014), and Goldman (2009). The aim of this paper is both to bury and praise it. RI faces serious objections (section 2). But from these objections I discern general desiderata on theories of normative reasons and construct an alternative theory which satisfies these desiderata (sections 3–4). I label the

[†] University of Nottingham, University Park, Nottingham NG7 2RD, UK; Email: neil.sinclair@nottingham.ac.uk

alternative 'Normative Reason Statement Purposivism' or NSRP, and its core claim is that a function of the making of reason statements is to create, via the particular deliberative mechanism of normative discussion, an appropriate motivation in the target agent. To help motivate and elucidate this theory, it is first necessary to examine RI.

Several elements of RI need explication. First, a normative reason to Φ is a consideration in favour of Φing, one that makes Φing to some degree justified, right, obligatory or 'to-be-done' (Scanlon 1997, 19). Second, 'existing motivations' refers to the elements of an agent's subjective motivational set (Williams 1981, 102). These elements may include desires and goals but also "dispositions of evaluation, patterns of emotional reaction, personal loyalties and various projects … embodying commitments of the agent" (Williams 1981, 102). One way of unifying such states (and avoiding triviality, cf. Korsgaard 1986) is by a common *mind-to-world* direction-of-fit: they are psychological states that aim, through prompting action in consort with relevant beliefs, to impose themselves on the world so as to have their content realised (Humberstone 1992). Third, to say that A in C can rationally access a motivational state from existing motivations is to say that A in C would access that state via a process of sound practical deliberation from existing motivations. (In what follows, I omit the qualification 'from existing motivations' when not salient.) Crucially, Reason Internalists hold that the outcomes of sound practical deliberation are constrained by the initial motivational set from which agents deliberate (Lillehammer 2000, 507–8). Finally, I assume that RI carries with it an implicit explanatory direction from facts about motivation to facts about reasons. On this interpretation, where A cannot rationally access the motivation to Φ, it is *because* of this fact that A has no reason to Φ (Hurley 2001).

Despite appearances, RI is a philosophical hydra, the myriad versions of which can be distinguished across at least three dimensions. First, regarding the account of 'sound practical deliberation'. According to Williams, for example, sound practical deliberation involves practical reasoning, concerning circumstances C, from a motivational set that has undergone a purge of relevantly false beliefs. A paradigm case of such reasoning is "that leading to the conclusion that one has reason to Φ because Φing would be the most convenient, economical, pleasant etc. way of satisfying some element in [the agent's subjective motivational set]" (1981, 104). Other Reason Internalists add further conditions, such as an initial process of establishing maximal coherence among a set of motivations (Smith 1995, 114–116; Goldman 2009, 57–82) and/or sensitivity to the normative pressures of others (Bedke 2010, 50–51).

Second, versions of RI can be distinguished according to the nature of the motivational state accessed. On strong versions, this is the action of Φing or the intention to Φ (Williams 1989, 35; Shafer-Landau 2003, 170; Bedke 2010, 39). On weaker versions, it is the state of having some desire or inclination to Φ,

where this state need not result in Φing or intending to Φ (Smith 1995, 110; Setiya 2014, 222).

Finally, Reasons Internalists differ regarding what it is to 'access' the relevant motivational state. An 'ideal' or 'fully rational' version of an agent A – call her A+ – is one who, faced with the same circumstances, undergoes a process of sound practical deliberation from existing motivations (Smith 1995, 112–116; Bedke 2010, 41). On a simple 'exemplar' model to say that A in C can access a desire (or intention) to Φ (or Φing) via a process of sound practical deliberation is just to say that A+ would possess that desire (or intention, or perform that action; Williams 1989, 35). On a more complex 'advisor' view, to say that A in C can access a desire (or intention) to Φ (or Φing) via a process of sound practical deliberation is to say that A+ would want A, as she was in the original circumstances, to possess that desire (or intention, or perform that action; Smith 1995, 112). In contrast, according to the 'imaginative projection' account, to say that A in C can access a motive (or intention) to Φ (or the act of Φing) via a process of sound practical deliberation is to say that A+ would want for herself to have that motive (or intention, or perform that action) were she in A's actual circumstances (Bedke 2010, 43).

2. *Relevance, attractions, detractions*

Though it glosses much detail, the foregoing is enough to help discern the relevance, attractions and detractions of Reasons Internalism.

First, relevance. Quite apart from its acting as a substantive constraint on what reasons there are, Reasons Internalism is one part of an attractive yet seemingly inconsistent triad (Brink 1997). The two other elements are:

> *Moral Rationalism*. Necessarily, if A in C is morally required to Φ, then A in C has normative reason to Φ.

> *Moral Absolutism*. Moral requirements are not contingent upon the actual motivational set of the agents to whom they apply.

If Reasons Internalism and Moral Rationalism are true, then an agent can be morally required to Φ only if that agent can rationally access the state of being motivated to Φ from existing motivations. Hence the contents of moral requirements on a given agent are contingent upon her existing motivations (Harman 1975). Not only is this contingency troubling in itself, it seems to conflict with the basic tenets of moral realism (Shafer-Landau 2003, 170). Similarly, if Reasons Internalism and Moral Absolutism are true then normative reasons are, whereas

moral requirements are not, contingent on the motivational set of the agents to whom they apply, and hence moral requirements cannot entail normative reasons (Foot 1972; Williams 1981, 110). There follows a surprising result concerning the authority of morality: moral requirements need not be something that we have reason to follow (Shafer-Landau 2003, 192–193). Finally, if Moral Rationalism and Moral Absolutism are true then at least some reasons, viz., those that accompany moral requirements, are not contingent on the actual motivational sets of agents.

This last option might seem the most appealing were it not for certain considerations which seem to favour Reasons Internalism. No doubt one motivation for accepting Reasons Internalism is the desire to show that those who fail to act in accordance with their strongest reasons are in some way acting against their own motivationally enshrined interests – a desire to hit them where it hurts, as Foot puts it (1978, 152). But as an argument this is liable to backfire, since accepting Reasons Internalism might equally warrant the conclusion that agents have fewer reasons than was previously thought, precisely because many of these putative reasons do not relate in the right sort of way to agents' existing concerns. A stronger argument can be mounted on the basis of an apparent conceptual connection between reasons and practical deliberation, that is, reasoning about what to do and feel. Many have suggested that one function of reasons is to guide action by providing considerations that agents can deliberatively cognise: reasons are for reasoning with. As Wong puts it: "what point could a reason have if it is not capable of motivating the agent who has it?" (2006, 540; cf. Lillehammer 2003, 42). Reasons Internalism is one way of capturing the required connection here, since rationally accessing the state of being motivated to Φ requires the relevant idealised agent to deliberatively cognise the consideration presented by the relevant reason. Note, however, that whilst this line of thought suggests a connection between reasons and deliberation, Reasons Internalism adds the further claim that deliberation proceeds from, and is constrained by, existing motivational sets. To secure this result, one can add the plausible thought that practical deliberation requires motivational input – some salient goal or end to reasons towards – and that only existing elements of agents' subjective motivational sets can provide that input (Williams 1981, 109; Milgram 1996, 198). These claims granted, Reasons Internalism is supported insofar as it provides a plausible explanation of the generally acknowledged connection between reasons and practical deliberation.

Yet Reasons Internalism also has detractions. First, it seems vulnerable to counterexample insofar as we can conceive of cases where agents have normative reason to Φ and yet they cannot rationally access the state of being motivated to Φ. Cases of conscientious objectors (Williams 1981, 106), unrepentant wife-beaters (Wong 2006, 540), neglectful parents (Schroeder 2007, 103) and indifferent

stairwell dwellers (Bedke 2010, 53) all fit this description. A recent example comes from Shafer-Landau (2009, 190):

> Consider an experienced torturer working on behalf of an authoritarian government. Such a person not only endorses the legitimacy of the regime, but takes active pleasure in breaking his victims. His greatest joy is stripping the last vestiges of dignity from those who initially resist his demands. At a given session, as he is about to apply the electrodes, he pauses to consider the merits of his action. He sees that doing so will get him what he most wants, and will frustrate none of his desires. He proceeds accordingly.

Here it seems that the torturer has a reason not to apply the electrodes and yet no amount of sound practical deliberation can lead him to access a motivation to refrain. (This last is contentious, of course, given the chronic under-description of both the torturer's psychology and the process of sound practical deliberation. But I put this to one side for now.) Hence, it seems, Reasons Internalism is false.

Some might worry that this argument is vulnerable insofar as it relies on intuitive judgements about reasons that beg the question (Shafer-Landau 2009, 192) or are not universally shared (Lillehammer 2000, 508) or reflect views on what it is *expedient* to say about reasons, rather than what it is *correct* to say (Williams 1981, 111; 1989, 40). But whatever its other failures, the objection here need not rely solely on intuition, for the judgements in these cases can in turn be explained in terms of another plausible claim about the function of normative reasons. This is the claim that part of the function of such reasons is to provide a degree of critical distance from the existing motivational states of agents. The thought is that a function of reasons is not just to reflect the ways in which agents are already incentivised to act, in virtue of their subjective motivational set, but to provide considerations that can form the basis of criticism of that set; criticisms that point the agent in new and better directions. As McDowell puts it: "Reason-giving explanations require a conception of how things ideally would be, sufficiently independent of how any individual's psychological economy operates to serve as the basis for a critical assessment of it" (1995, 76; cf. Thomas 2006, 86; Wong 2006, 537). It is just this independence that is displayed in our intuitive judgements about the reasons of torturers and the like.

Reasons Internalists are aware of this problem and it is in fact partly because of it that they accept a view which moves beyond the sub-Humean model. Reasons Internalism does not constrain agents' reasons by their motivations as they stand but by the motivations they can access after sound practical deliberation. Such a process can involve correcting for errors of fact and mistakes of reasoning. Accordingly Reasons Internalism can allow some critical distance between the agents' reasons and their *current* motivational sets. As Williams notes, "this is … enough for the notion [of a reason] to be normative" (1989, 36).

However, one might worry that this manoeuvre fails to provide *sufficient* critical distance (McDowell 1995, 77; Wong 2006, 551). According to the current view, there is no scope for normative reasons to provide critical perspectives on sets of motivations that agents would possess after undergoing sound practical deliberation from existing motivations. It is plausible that a function of reasons is to provide precisely this type of normativity, that is, to "provide a basis from which to criticize … desire *on the most fundamental levels*" (Wong 2006, 551). It is this phenomenon that is arguably on display in our judgements concerning agents such as Shafer-Landau's torturer, for though they are specified in ways which suggest they have engaged in sound practical deliberation, there is an obvious sense in which we can be critical of them, and critical in a way that appeals to considerations for and against actions – reasons – which they ignore (Shafer-Landau 2009, 190–191, 201). It follows not that all normative reasons reach beyond the motivations that can be accessed after sound practical deliberation, but that some do. This is nevertheless sufficient to refute Reasons Internalism.

These considerations generate an impasse. On the one hand, the idea that a function of reasons is to provide inputs into practical deliberation provides abductive support for Reasons Internalism. On the other, the idea that a function of reasons is to provide a critical perspective on motivational sets suggests that Reasons Internalism is false. Rather than take sides in this debate, I propose a concessive strategy. Each of the above arguments can be understood as drawing plausibility from an implicitly understood desideratum on successful theories of normative reasons. My strategy is to tease out these desiderata and consider whether a single theory can satisfy both.

First, then, consider the argument in favour of Reasons Internalism. This draws its plausibility from the following (pro tanto) desideratum:

> D1. A theory of reasons should accommodate the thought that one function of reasons is to provide inputs into practical deliberation.

The argument against Reasons Internalism suggests a distinct desideratum:

> D2. A theory of reasons should accommodate the thought that one function of reasons is to provide critical distance from agents' motivations (both as they are now, and as they would be after sound practical deliberation).

In fact, these desiderata can be understood in a complementary way, for D1 provides a partial specification of the mechanism whereby the mode of criticism referred to in D2 might come to affect agents' motivations. It is important, though, not to misunderstand the strength of these desiderata. The arguments in their favour do not support the view that these claims exhaust the functions of

reasons. The case in support of D2 does not support the view that *every* reason is part of a critical perspective distanced from agents' actual motivations. Likewise the argument that supports D1 does not support the claim that *every* reason is an input into practical deliberation (after all, reasons can go unrecognised). But, so understood, D1 and D2 seem to capture two plausible thoughts about the nature of reasons, at least as reflected in existing debates.

3. *Williams and normative discussion*

In sections 4 and 5 I will put forward a theory of reasons – NSRP – that seems to capture what is plausible about D1 and D2. One way to approach this theory is by first considering Williams' famous argument for Reasons Internalism. Though there is considerable controversy regarding the interpretation of Williams' views (e.g. Thomas 2006, 69–81; Finlay 2009), a brief detour through some of these debates will be helpful to elaborate the key notion of normative discussion which features in NSRP.

Williams uses the phrase 'external reason' to refer to any (putative) reason that needn't (in order to be a reason) satisfy the necessary condition laid down by Reasons Internalism and 'internal reason' to refer to any (putative) reason that does need (in order to be a reason) to satisfy this condition (cf. Finlay 2009, 17). Thus the debate about Reasons Internalism can be understood as the debate about the existence of external reasons. On the traditional yet controversial interpretation of Williams' argument against external reasons, there are two key premises.[1] The first is the so-called explanatory constraint: that a putative reason is a reason for an agent only if that agent would, if ideally rational, be appropriately motivated by believing in that reason (Williams 1981, 106; 1989, 39; cf. Thomas 2006, 69–70; Finlay 2009, 3–4). The second is the claim that an agent can, if ideally rational, be appropriately motivated by believing in a reason only if she antecedently has some motivational state that could bring it about that she be so motivated (Williams 1981, 109; cf. Finlay 2009, 4). On this interpretation, Williams' second premise presents a Humean instrumentalist conception of rationality: the view that "practical reason cannot by itself produce or entail a motive, but merely facilitates flow of motivational force from pre-existing desires, channelling it from one object to another" (Finlay 2009, 4).

So understood, Williams' argument has multiple detractors. Many have pointed out that it is question-begging to assume an instrumentalist conception of rationality in the course of an argument against external reasons (Finlay 2009, 4), as the defender of external reasons will prefer a substantive theory according to which an agent can be rational so long as she appropriately responds to her reasons,

[1] For examples of the traditional interpretation, see Korsgaard (1986), Hooker (1987), Smith (1995), and Shafer-Landau (2003, 172). For controversy, see Thomas (2006, 69–81), Finlay (2009).

regardless of such reasons' instrumental connection to an initial motivational set. So, for example, a previously callous agent who comes, via a process of non-deliberative *conversion*, to be appropriately motivated by a belief about the reason-giving nature of others' pain, can be rational even if, pre-conversion, she had no desire-like state that could ground this motivation (McDowell 1995, 74). Other critics of Williams' apparent instrumentalism attempt to provide more detail about the sorts of non-instrumentalist processes that might be involved in rational deliberation. Milgram, for example, emphasises the role of 'tutelage of experience'. He elaborates:

> … just as reasoning can be practical as well as theoretical, so experience can be practical too. Using 'fact' and 'value' for a moment as contrasting terms, we can say that there may be no relevant *fact* about the upcoming situation of which I am unaware, and there may be no element of my [subjective motivational set] that I am unable to deploy deliberatively, but this does not mean that I can know what it will be *like* living through it. One reason for this is that what something is like is, often enough and in large part, a matter of my evaluation of it; and it is this evaluative aspect of what it is like that makes what it is like not simply a further fact available to the internalist. (1996, 207)

For example:

> Imagine someone brought up on the pleasures of light verse … We might call a strong appreciation for the accomplishments of lesser poems … a component of his [subjective motivational set]. Now suppose he is directed to a passage such as, say, the second part of Yeats' 'A Dialogue of the Self and Soul'. It is characteristic of such poems that appropriate reactions to them are complex and not easily summarized, but we may suppose that one component of his reaction amounts to, 'I had no idea a poem could be like *that*', and that his [subjective motivational set] now contains desires to read more poetry of this kind. (1996, 211)

One issue arising is whether such experiential tutelage is distinct from 'sound practical deliberation' of the sort Williams (on the instrumentalist interpretation) countenances. Williams takes the latter notion to be somewhat indeterminate, with "no fixed boundaries on the continuum from rational thought to inspiration and conversion" (1981, 110). Nevertheless, as Milgram (1996, 210–212) notes, it is suggestive of a Humean interpretation that all of Williams' examples of sound practical reasoning are instrumental, that is, aimed at finding ways of satisfying (or promoting) elements of one's existing motivational set. In the poetry example, the derivation of new motivations does not seem to be instrumental in this way (for a further example, see Wong 2006, 539 ff.)

Are there further examples or non-instrumental rationality? One striking fact about the types of rational deliberation so far discussed is that they are all essentially *private*, that is, they describe processes that agents go through without direct

interaction with other agents. Williams' agents consider the best way to satisfy antecedent desires; Milgram's agent reads a poem. But, face-to-face conversing with other agents about what to do, what ought to be done and what reasons there are in the case at hand is a typical, perhaps *the* typical, way in which agents come to revise their views about their reasons. Consider the following example.

> Sally, Elke and Luna are members of an academic department deciding whether or not to institute a policy of anonymous marking. Sally thinks anonymous marking really ought to be implemented, and cites in support the possibility of implicit biases that can affect non-anonymous marking. Elke disagrees, claiming: "We ought not to do this, it would undermine a vital type of rapport that is essential in any good teacher–student relationship". Luna is initially undecided, but notes that "The new system would have huge administrative costs, and we should think carefully about the effect of extra workload on overworked administrators." After much discussion they decide that they have most reason to implement anonymous marking.

Sally, Elke and Luna are engaged in an activity that can be labelled 'normative discussion' (cf. Gibbard 1990). Such discussion aims at consensus regarding what to do or feel and involves as a crucial part agents citing putative reasons (e.g. "this would undermine rapport") in the hope or expectation that other agents will come to be moved by the considerations therein presented. In some, but not all, cases, this will result in a change in motivational profile consequent on (or simultaneous with) the acquisition of a new reason-belief.

If we suppose the case described to be at all representative of the actual, messy, ways in which agents collectively deliberate about what to do, it is plausible that some of the motivational shifts that result are instances of rational deliberation. Consider Elke, who is not initially motivated in favour of anonymous marking. After the discussion she comes to believe that she has a reason to support anonymous marking. Her case can plausibly be described as 'considering the matter aright' (Williams 1981, 108–109) via a respectable mechanism, namely the mechanism of reflectively engaging in normative discussion. This mechanism also seems to be distinct from instrumental deliberation from existing motivations – for on at least one way filling in the details of this scenario, it is not that Elke now supports anonymous marking because she has come to believe that it is in fact the best way to promote her antecedent goal of teacher–student rapport; rather, the interaction with her colleagues has caused Elke to reassess he priorities and acquire new non-instrumentally derived motivations, so that she now places teacher–student rapport as only one concern among others. So described, we seem to have another – socialised – counterexample to instrumentalism.

The idea of mechanisms whereby agents' motivations are directly sensitive to the normative pressures of others is familiar from other contexts. As Bedke (2010, 51) notes, the experiments of Milgram (1974) and others suggest that humans possess motivational mechanisms that "take certain pressures from others

as input and output novel motivations or motivational strengths". Such mechanisms also have a key role in Haidt and Bjorklund's empirically grounded theory of moral judgement, which postulates a 'reasoned persuasion link', whereby moral judgements are used to persuade others to share the motivational tendencies of the speaker and "rhetoric is the art of pushing the ever-evaluating mind over to the side the speaker wants it to be on" (2008, 192).

This mention of compliance with authority and rhetoric might ring alarm bells. Williams explicitly rules out being "persuaded by … moving rhetoric" (1981, 108) as an acceptable, rational, way of coming to be moved by one's reasons. If the type of motivational influence present in normative discussion is no more than being persuaded by rhetoric or blindly following the demands of an authority, then Williams might rightly argue that this is not an instance of practical rationality at all. But it is useful to distinguish here. Rhetoric and demands from authority have two salient elements. First, they are most commonly parts of attempts to change the motivational profiles of others. Second, their method is in an important sense 'non-deliberative'. That is, they hope to affect this change not by drawing attention to pre-existing features of the situation, but, in the case of rhetoric, by using emotive language intended to directly elicit the required reaction and in the case of demands from authority by making the demand itself the reason for compliance ('Because he told me to'). In this way, both are examples of what Falk (1953) labels 'goading'. What Falk contrastively labels 'guiding' shares the first but not the second feature. In guiding, agents attempt to change the motivations of others by drawing their attention to features of the situation that exist independently of any demand being made, in the hope that the target agent will recognise their normative and hence motivational salience (i.e. come to judge that they are reason-giving and be motivated accordingly). In guiding, the persuasive force which the utterance has relies on what it says, rather than on simply being the making of a demand. On this taxonomy, the type of motivational influence in play in normative discussion is paradigmatically guiding rather than goading (although no doubt actual normative discussion typically involves both types of influence). That is to say, normative discussion seeks to alter agents' motivations in part by adducing independent considerations that can subsequently be cited as reasons in support of those motivations.

This point allows a response to the problem raised in the previous paragraph. The comparison with rhetoric and adherence to authority is useful insofar as it demonstrates the empirical possibility of a type of mechanism for generating motivations that goes beyond instrumental deliberation from existing motivations. Most generally, these are mechanisms which take contributions from others as inputs and output novel (i.e. non-instrumentally connected) motivations. But Bedke and others seem over-hasty to claim that the very same mechanisms at play in rhetoric and adherence to authority are at play in normative discussion. Both types of

mechanism allow for the contributions of others to generate novel motivations, but the latter involves guiding rather than (mere) goading.

All of this spells trouble for the defender of Reasons Internalism who takes what is 'rationally accessible' to be limited to what is accessible by instrumentalist reasoning from existing motivations. Such a Humean conception of a 'sound deliberative route' is put under severe pressure by these apparently non-instrumentalist ways of 'rationally accessing' relevant motivation.

How might Williams, or his defenders, respond? One option is to cry foul on the point of interpretation. As previously noted, it is a matter of considerable contemporary controversy whether Williams did in fact base his case against external reasons on instrumentalism. Finlay, for example, has given persuasive arguments for thinking that Williams' argument is based rather on the point that the concept of a (practical) reason just is the concept of an explanation of action under the condition of sound reasoning (Finlay 2009, 13–19; cf. Thomas 2006, 69–75). On this account, Williams would insist that the reasons we have are constrained by what is accessible by sound practical deliberation from where we begin, but would adopt a more expansive conception of such deliberation that could include, for example, the type of reasoning involved in the case of Elke et al.

Issues of interpreting Williams aside, such a version of Reasons Internalism has recently been defended by Bedke. The key to this position is the claim that the persuasive mechanisms involved in normative discussion are included as instances of practical deliberation that define the limits of rational accessibility. More precisely, on Bedke's view, A has reason to Φ only if A+ would want for herself to Φ were she in A's actual circumstances (2010, 43–44). Crucially, the norms of rationality that A+ internalises are not purely instrumental, for they include the norm of being sensitive to the legitimate normative pressures of others, where this normative pressure is precisely the kind of pressure exerted in the 'guiding' sort of discussion described above (2010, 48–53).

Such a move may seem ad hoc, for what justification could there be to include following this norm of 'social rationality' in the account of what constitutes ideal rationality? Bedke's answer is the 'promotion account', which holds that a norm is a norm of rationality "because, when embodied in one's psychology, it promotes the objects of one's motivations, or more narrowly ends, in the kinds of circumstances in which the agent typically finds himself" (2010, 48). This, Bedke claims, explains why the types of deliberation Williams focuses on count as following norms of rationality (2010, 49). But it also explains why a degree of sensitivity to the type of influence that occurs in normative discussion counts as following a norm of rationality: "By participating in certain normative pressures, the objects of antecedent motivations are likely to be served, whatever they happen to be, for the objects of most motivations cannot be served at all, or easily served, without a supportive social environment" (2010, 52).

This socialised Reasons Internalism certainly nullifies the case of Elke et al. as a potential counterexample. But is it plausible? Bedke argues that it goes some way to dealing with the problem generated by apparently external reasons, such as the torturer's reason to stop torturing (section 2). He considers the case of Andy, "who doesn't care about strangers in a burning building as he finds himself alone in a stairwell next to a fire alarm":

> … imagine that Andy+ is somewhat sensitive to the normative pressures of others. The suggestion … is that Andy+ is not fully ideally rational unless he includes the idealization of the norms by which Andy is sensitive to the legitimate normative pressures of others ... If this dimension of rationality is included, there will likely be an ideal version of Andy who would want for himself to pull the fire alarm were he in Andy's shoes, viz. a version of Andy that has been hypothetically subjected to normative pressures to save the strangers in the building. (2010, 53)

Thus this version of Reasons Internalism is consistent with the claim that Andy has a reason to pull the alarm.

On reflection, however, problems arise. We are asked to imagine that Andy+ is hypothetically subjected to the (legitimate) normative pressures of others. But which others? If we suppose that the relevant others are those representing the interests of the people in the burning building, then it seems likely that Andy+ would want for himself to pull the alarm, were he in Andy's shoes. But if we suppose the relevant others to be some totalitarian group (of which Andy is an enthusiastic member) dedicated to the eradication of the type of people occupying the building, it seems just as likely that Andy+ would not want for himself to pull the alarm, were he in Andy's shoes. Which hypothetical normative pressures are relevant to Andy's reasons? Note that one thing that Bedke cannot say is that the relevant pressures are those applied by people who correctly perceive the reasons in the case. For, as he notes (2010, 45), if one wishes to employ Reasons Internalism as a non-trivial constraint on reasons, one cannot rely on a conception of rationality according to which what counts as rational is simply a matter of responding appropriately to reasons.

More plausibly then, Bedke could argue that the relevant others are those individuals present in the social environment in which Andy (typically) finds himself. On this view, we hypothetically subject Andy+ to the normative pressure of agents taken from Andy's milieu. This would include the normative pressure of the agents in the burning building. It seems right that if these are the *only* agents hypothetically subjecting Andy+ to normative pressure, then Andy+ would want himself to pull the alarm, were he in Andy's shoes. But suppose, as before, that Andy's actual situation is that he (contentedly) lives in a totalitarian state dominated by a group whose explicit goals involve the eradication of people such as

those in the burning building. On the current account, the normative pressures of these agents are included in the relevant hypothetical scenario, and, given certain assumptions about the pervasiveness and stability of the totalitarian state, it seems likely that in this case, Andy+ would want for himself not to pull the alarm, were he in Andy's shoes. Yet, intuitively, even in this case, Andy still has a reason to pull the alarm (suppose that doing so would activate the building's sprinklers, for example). This case therefore seems to be a counterexample even to the more inclusive version of Reasons Internalism that Bedke prefers.

Can we offer a diagnosis? The socialised view seems promising insofar as it recognises a connection between reasons and the deployment of those reasons in normative discussion. But in accepting the promotion account of rationality, the view also explicitly takes an instrumentalist flavour – for on that account, engagement in normative discussion is ultimately justified in virtue of its propensity to better satisfy one's antecedent motivations (in a social environment). The problem arises because even in the social environment an agent faces, these antecedent desires might best be served by being motivated in ways in which one intuitively has no reason to be motivated. The social view still fails to provide the requisite critical distance.

4. Normative Reason Statement Purposivism

The foregoing emphasises a tight connection between reasons and the process of giving and asking for reasons in normative discussion. But it seems that incorporating this thought into the formulation of Reasons Internalism is unsuccessful. It is worthwhile, therefore, to consider the possibility of alternatives that build on the insight without falling prey to the same mistakes.

In her seminal paper, Foot remarks that "I do not understand the idea of a reason for acting, and I wonder whether anyone else does either" (1978, 156). Earlier, I expressed D1 and D2 in terms of the 'functions' of reasons. This seems to compound Foot's problem: if reasons are obscure, their functions are doubly-so. But we can avoid both worries by switching focus from *reasons* to *statements of reasons* (e.g. Williams 1981, 101; Thomas 2006, 69), and take our intuitions in cases such as Shafer-Landau's torturer to concern not the existence of reasons, but the appropriateness of making reason statements. Let us say that to make a normative reason statement is to sincerely utter a sentence of the form 'F is a reason for A to Φ' or 'A has a reason to Φ', where 'reason' is used normatively. In this light, we can recast D1 and D2 as follows:

> D1*. A theory of normative reason statements should accommodate the thought that a function of such statements is to provide inputs into practical deliberation.

> D2*. A theory of normative reason statements should accommodate the thought that a function of such statements is to provide critical distance from agents' motivations (both as they are now, and as they would be after sound practical deliberation).

We can now consider whether a theory can be constructed precisely to meet these desiderata. One good place to start is to consider how such statements are *used*, in particular, in the context of normative discussion. Consider the following:

> Normative Reason Statement Purposivism (NRSP): Necessarily, a function of reason statements of the form 'A has (normative) reason to Φ' (etc.) is to deliberatively create in A some motivation to Φ.

Here 'deliberatively create' means 'create through the guiding mechanisms partly constitutive of normative discussion', understanding normative discussion in the ways elaborated above. These mechanisms are ones of guiding insofar as they involve the presentation of independent considerations which, if things go well, bring forth in the target agent an appropriate motivation. This motivational shift can occur in one of at least two ways. In the first type of case, reason statements create motivational shifts by engaging with existing motivations, highlighting paths of actions which are incentivised in light of those motivations and citing features which help explain why this is so. In the second type of case, whose possibility is suggested by the previous account of normative discussion, reason statements do not engage with existing motivations in the same way, but do hope to be part of a guiding process whereby novel motivations arise.

NRSP also talks of a 'function' of normative reason statements. The sense of function here is semantic: that reason statements have this function is a feature of their conventionally enshrined meaning, in the same way that the meaning of imperatives, say, is partly defined by their function of producing compliance. According to NRSP, when a sincere speaker makes a normative reason statement, the conventions surrounding the use of such statements mean that those who understand the speaker take them to be attempting to deliberatively alter the target's motivational set. A natural corollary of NRSP is the further claim that normative reason statements sometimes fulfil this semantic function, in such a way that it is also an aetiological function, that is, an effect that such statements have had in the past which explains the continued proliferation of tokens of that type.

Is NRSP plausible? Consider D1* and D2*. According to NRSP, a function of reason statements is to deliberatively create new motives, and they fulfil this function, when they do, by providing inputs into the deliberation of target agents. Hence D1* is satisfied. Insofar as such statements hope to fulfil this function via the first type of 'guiding' mentioned above, there will be distance from the target

agents' current motivations; insofar as reason statements hope to fulfil this function via the second type of guidance mentioned above, there will be distance from the target agents' motivations even as they would be after sound practical deliberation. Hence D2* is satisfied. The fact that NRSP can capture these plausible thoughts about the functions of reason statements is one important argument in its favour.

5. *NRSP and the connection between reasons and motives*

Return to the original question: What is the connection between it being true that A has a reason to Φ and the possibility of A being motivated to Φ? Given the idiom-shift of the last section, the question becomes: What connection is there between the appropriateness of a statement of the form 'A has (normative) reason to Φ' (etc.) and the possibility of A being motivated to Φ?

NRSP assigns a semantic function to normative reason statements. If a particular type of reason statement (e.g. 'You have reason to leave') is to fulfil this function, some of its tokens must deliberatively create a relevant motive. It follows that if a particular type of reason statement cannot in this way deliberatively create this motive, then it cannot fulfil this function. Assuming that a statement that cannot fulfil one of its semantic functions is in that respect not semantically appropriate, it follows that where there is no possibility of a particular type of normative reason statement deliberatively creating the relevant motive, then that particular type of statement is, in one respect, semantically inappropriate. In other words, given these assumptions, NRSP entails the following claim (mildly reminiscent of Reasons Internalism):

> RI-ish: Necessarily, statements of the form 'A has (normative) reason to Φ' (etc.) are in one sense semantically appropriate only if the making of such statements can deliberatively create in A some motivation to Φ.[2]

Here 'in one sense semantically appropriate' means 'appropriate in light of the semantic function assigned by NRSP'. According to RI-ish there is a necessary connection between a type of normative reason statement being in some sense semantically appropriate and the (historical and nomological) possibility of target agents being relevantly motivated by statements of that very type. Consider, for example, Shafer-Landau's torturer. Suppose that it is simply not possible (given the history of the world and the laws of nature) that the torturer come to be

[2] I take this view to have close affinities with the version of Reasons Internalism defended by Manne (2014).

motivated to stop torturing on the basis of other agents exerting normative pressures on him, for example by making reason statements such as: 'The pain you're causing is a very good reason to stop torturing, you know.' Then, according to RI-ish, there is at least one sense in which such statements are not semantically appropriate (note that other types of evaluative statements may be available – see next paragraph). At this point, we may need to deal with the torturer in ways other than engaging him in normative discussion. Williams (1981, 43) labels such characters "hopeless or dangerous".

Crucially, in comparison with Williams' version of Reasons Internalism, RI-ish extends the 'critical distance' between agents' existing motivational sets and their reasons in two key ways. First and most importantly, it allows that agents' reasons are not in any way constrained by their initial motivational sets. Although Williams' Reasons Internalism does not limit reasons to initial motivations, it does limit reasons to the outcomes of sound practical deliberation, and those outcomes *are constrained by the agent's initial motivational set* (see sections 1 and 3). RI-ish, on the other hand, limits agents' reasons only to what can be reached after engagement with normative discussion – there is no requirement that the outcomes of *that* process be restricted by initial motivations. Second, by explicitly including interpersonal normative discussion in the types of processes that might be involved in sound practical deliberation, RI-ish provides more critical distance than Williams' view even on the assumption (which it rejects) that the outcomes of sound practical deliberation *are* constrained by existing motivations. For, as previously noted (section 3), Williams' view of sound practical deliberation is distinctively *private* – it involves no interpersonal normative discussion. As it is plausible that one can reach *further* from initial motivations by a process that includes interpersonal normative discussion than by a process that excludes it, including this process in the mechanisms of sound practical deliberation already increases critical distance between existing motivations and reasons (this was one of the merits of Bedke's approach, which unfortunately failed on other grounds). Combining these two points, the key point is that unlike all other versions of Reasons Internalism, RI-ish does *not* tie agents' reasons to their existing motivational set – rather it ties agents' reasons to their deliberative possibilities (and has a liberal, interpersonal, view of deliberative possibilities). Finally, note that neither Williams' view nor RI-ish are themselves committed to Moral Rationalism (section 2). If they reject it, then they can insist that other evaluative judgements (e.g. of wickedness or depravity) are entirely appropriate of characters like the torturer. For such terms, critical distance is as great as any party to the debate considers reasonable. On this view, it is only judgements of reasons which are (somewhat) more restrictive, but this very restriction is a natural upshot of their particular functional role, as given by NSRP.

Some important caveats remain. First, I assumed above that if a particular type of statement *cannot* fulfil a particular (semantic) function, then it is in that respect (semantically) inappropriate. But it is not plausible to assume that where a particular statement *does not* fulfil a particular function, then it is in the same way inappropriate. For example, one function of commands is to produce compliance, but just because a command is not complied with, doesn't mean it was (with respect to this function) inappropriate. Likewise, just because a particular reason statement does not fulfil its function of deliberatively creating a relevant motive does not mean that that reason statement was (with respect to this function) inappropriate. Consider again Shafer-Landau's torturer. We saw that if there is no possibility of him being motivated by normative reason statements, then in one sense those statements are inappropriate. But suppose that there is *some* possibility of him being motivated by such statements, but that after making them on a particular occasion, this possibility is not realised. It does not follow that the statements were inappropriate. It may be that his interlocutors were justified in engaging him in normative discussion insofar as such discussion is a less costly method of securing the hoped-for social co-ordination (compared, for example, to coercion, brainwashing or physical restraint), and other things being equal, one is justified in deploying less costly methods first. On this view, so long as there was at least a possibility of these reason statements deliberatively creating the relevant motive, their deployment can still be appropriate (with respect to their NRSP function). Similarly, even if there is no possibility of the relevant motive being deliberatively created, but there is some uncertainty about whether this is so, the low cost of reason statements in comparison with alternatives can make their employment appropriate.

Note, therefore, that on this account there will be some indeterminacy whether particular reason statements are semantically appropriate (with respect to their NRSP function). In some cases it may be unclear whether deliberatively creating the relevant motive is possible. In other cases, it may be clear that this is possible, but unclear how likely it is. These uncertainties will make (this dimension of) the semantic appropriateness of the reason statements similarly uncertain. Yet this is a positive feature of NRSP, since it reflects experience concerning reason statements. As Williams puts it, there is some indeterminacy in the way in which 'the presence of deliberative reasons … *fall off* in one or another direction' (1989, 43). For example, it is often worthwhile engaging in normative discussion with an adult human, but seldom with an obdurate donkey. Yet we should not fool ourselves by denying that these two exist on a continuous scale of susceptibility to reasons. Discerning precisely where and how normative discussion – and the reason statements it deploys – becomes redundant is not an easy matter.

A second caveat is that RI-ish relates only to one function of normative reason statements and leaves open the possibility that such statements have

complementary functions. With respect to these, it might be that a particular type of statement is appropriate even if there is no possibility of its tokens deliberatively creating the relevant motivational state. For example, an additional function of normative reason statements may be to signal to third parties what sort of motivational states one is generally disposed to encourage or discourage. With respect to this signalling function, it might be that particular types of reason statements are appropriate even when there is no possibility of their tokens deliberatively creating the relevant motivational state.

RI-ish therefore provides a nuanced picture which goes some way to explaining somewhat conflicting thoughts about cases like Shafer-Landau's torturer. If the torturer is immune to persuasion via normative discussion, there is one sense in which statements to the effect that he has (normative) reason to stop are inappropriate – namely the sense in which they cannot fulfil their function of deliberatively creating the relevant motivational state. However, if it is unclear whether the torturer is immune to such persuasion, such statements can be worthwhile insofar as they represent relatively low-cost attempts to induce the relevant state. Finally, regardless of their possible deliberative effect on the torturer, such statements may be appropriate with respect to other, for example signalling, semantic functions.

We can make all these points more vivid by imagining a confrontation with an ardent opponent of Reasons Internalism. The initial problem, this opponent reminds us, is that Reasons Internalism does not generate sufficient critical distance between an agent's actual motivations and her reasons. This is perfectly exemplified by the sadistic torturer, who has reasons to stop torturing, yet cannot come to be appropriately motivated. But, our opponent continues, the move from Reasons Internalism to RI-ish doesn't help increase this critical distance at all. For, according to RI-ish, if it is the case that we cannot, through the mechanism of normative discussion, create in the torturer some motive to stop, then it is not semantically appropriate to assert that he has a reason to stop. But this is (surely!) an appropriate thing to assert. So RI-ish is no better than RI, at least when it comes to providing appropriate critical distance. To which the defender of RI-ish can reply as follows. First, we may not be in a position to assuredly know that there is no possibility of the appropriate motive being deliberatively created in the torturer, especially if we do not attempt to deliberatively engage him. If we do not know this, then making the reason statement may still be semantically appropriate (indeed, it might be appropriate precisely as part of an attempt to ascertain whether the relevant possibility holds). Second RI-ish captures only *one* dimension of the semantic appropriateness of reason statements. It is possible that such statements perform other functions (such as signalling), such that, by reference to these functions, the reason statement about the torturer's reason to stop is appropriate. To put it briefly if imprecisely: even if the reason statement cannot move the torturer, it

may be appropriate insofar as it has the potential to move (or reinforce the movements of) others. It is at this point that RI-ish appears more nuanced than the opponents' position: for it can capture the feeling that although in one sense the reason statement directed at the torturer is apt, there is another sense in which it is – for the purposes of rational debate – out of place or futile (after all: why reason with a donkey?). The mixed reactions here can be traced back to the multiple functions of such statements.

Before concluding, it is worthwhile to consider one potential objection to NSRP and is corollary RI-ish.[3] The objection is targeted at the claim that the function of reason statements as given by NSRP (viz., that of deliberatively creating an appropriate motive) is part of the conventionally enshrined meaning of such statements. The objection is that this cannot be the whole story of their meaning, because reason statements can also appear unasserted in embedded contexts (such as in the antecedents of conditionals), where they clearly do not have this function. This recalls the famous Frege–Geach problem for understanding unasserted normative sentences (cf. Schroeder 2008). In response, it is important to note that NSRP is not intended to be a *complete* story of the meaning of reason statements of the form 'A has (normative) reason to Φ'. The claim that the function of such statements is to deliberatively create motives is compatible with a number of (distinct) views about precisely how this function is realised. It is consistent with NSRP, for example, that reason statements function semantically to offer some description of a worldly relation (e.g., that A's Φing would promote some salient end; cf. Finlay 2014). In this case NSRP would simply add that the description is offered for the purpose of deliberatively creating relevant motivation. On this view, the descriptive or representational content of reason statements is the uncontroversial basis of explaining their meaning when embedded. Yet it is also consistent with NSRP that reason statements possess their NSRP-ascribed function by being expressive of particular non-cognitive attitudes. In this case, providing an account of the meaning of reason sentences in embedded contexts will be a matter of adapting the (controversial) general accounts of how expressive meaning can embed (e.g. Gibbard 1990; Blackburn 1998, 68–77; cf. Schroeder 2008). In the present context, however, this compatibility with both cognitivist and expressivist accounts of the wider meaning of reason statements is an advantage of NSRP. Insofar as the task with which we began was to give an account of reasons (or reason statements) which satisfied the two desiderata discernible from existing debates about Reasons Internalism, a theory which does exactly this while giving no further hostages to fortune is surely preferable. It seems that NSRP is such a theory.

[3] My thanks to an anonymous referee for pressing me on this point.

6. Conclusion

When it comes to characterising the connection between reasons and motives, Reasons Internalism is on shaky ground. Although it provides a plausible explanation of the thought that reasons are for reasoning with, it fails to allow for the requisite critical distance between some agents' reasons and their existing motives. This is so, I have argued, even in the case of the social type of Reasons Internalism suggested by Bedke. I have suggested that a theory that focuses instead on the (semantic) functions of normative reason statements goes some way to explaining the sometimes tenuous connection between the appropriateness of making such statements and the possibility of appropriate motivation in the target agent. According to that theory, reason statements cannot serve one of their functions (and hence cannot be in that sense appropriate) unless the target agent possesses some pre-existing propensity to be swayed by the making of them. This dimension of appropriateness therefore gives one sense in which 'reasons' are dependent on the motivational propensities (broadly construed) of target agents. But this is far from the sense suggested by Reasons Internalism.

Acknowledgments

This work was supported by the Arts and Humanities Research Council [grant number AH/J006394/1]. I would also like to thank three anonymous referees and Philipp Blum, acting for this journal, for extensive help in improving the paper from earlier drafts.

References

Bedke, M. 2010, "Rationalist Restrictions and External Reasons", *Philosophical Studies*, **151**, 1, pp. 39–57.
Blackburn, S. 1998, *Ruling Passions*, Oxford: Oxford University Press.
Brandt, R. 1979, *A Theory of the Good and the Right*, Oxford: Oxford University Press.
Brink, D. 1997, "Kantian Rationalism: Inescapability, Authority and Supremacy", in: B. Gaut and G. Cullity, eds, *Ethics and Practical Reason*, Oxford: Oxford University Press.
Falk, W. 1953, "Goading and Guiding", *Mind*, **62**, pp. 145–171.
Finlay, S. 2009, "The Obscurity of Internal Reasons", *Philosophers' Imprint*, **9**, 7, pp. 1–22.
Finlay, S. 2014, *Confusion of Tongues*, Oxford: Oxford University Press.
Foot, P. 1972, "Morality as a System of Hypothetical Imperatives", *Philosophical Review*, **81**, pp. 305–316.
Foot, P. 1978, "Reasons for Actions and Desires", in: P. Foot, *Virtues and Vices and Other Essays in Moral Philosophy*, Oxford: Oxford University Press.
Gibbard, A. 1990, *Wise Choice, Apt Feelings*, Harvard, MA: Harvard University Press.
Goldman, A. 2009, *Reasons from Within*, Oxford: Oxford University Press.
Haidt, J. and Bjorkland, F. 2008, "Social Intuitionists Answer Six Questions about Moral Psychology", in: W. Sinnott-Armstrong, ed., *Moral Psychology vol. 2*, Oxford: Oxford University Press.
Harman, G. 1975, "Moral Relativism Defended", *Philosophical Review*, **85**, pp. 3–22.
Hooker, B. 1987, "Williams' Argument Against External Reasons", *Analysis*, **47**, 1, pp. 42–44.

Humberstone, L. 1992, "Direction of Fit", *Mind*, **101**, pp. 59–83.
Hurley, S. 2001, "Reason and Motivation: The Wrong Distinction?", *Analysis*, **61**, 2, pp. 151–155.
Korsgaard, C. 1986, "Skepticism About Practical Reason", *Journal of Philosophy*, **83**, pp. 5–25.
Lillehammer, H. 2000, "The Doctrine of Internal Reasons", *Journal of Value Inquiry*, **34**, 4, pp. 507–516.
Lillehammer, H. 2003, "The Idea of a Normative Reason", in: P. Schaber and R. Huntelmann, eds, *Grundlagen der Ethik*, Frankfurt: Ontos Verlag, pp. 41–65.
Manne, K. 2014, "Internalism about Reasons: Sad But True?", *Philosophical Studies*, **167**, 1, pp. 89–117.
McDowell, J. 1995, "Might There Be External Reasons?", in: J. Altham and R. Harrison, eds, *World, Mind and Ethics*, Cambridge: Cambridge University Press.
Milgram, E. 1996, "Williams' Argument Against External Reasons", *Noûs*, **30**, 2, pp. 197–220.
Milgram, S. 1974, *Obedience to Authority: an Experimental View*, New York: Harper & Row.
Scanlon, T. 1997, *What We Owe To Each Other*, Harvard, MA: Harvard University Press.
Schroeder, M. 2007, *Slaves of the Passions*, Oxford: Oxford University Press.
Schroeder, M. 2008, *Being For*, Oxford: Oxford University Press.
Setiya, K. 2007, *Reasons Without Rationalism*, Princeton, NJ: Princeton University Press.
Setiya, K. 2014, "What is a Reason to Act?", *Philosophical Studies*, **167**, 2, pp. 221–235.
Shafer-Landau, R. 2003, *Moral Realism: A Defence*, Oxford: Oxford University Press.
Shafer-Landau, R. 2009, "A Defence of Categorical Reasons", *Proceedings of the Aristotelian Society*, **109**, 2, pp. 189–206.
Smith, M. 1995, "Internal Reasons", *Philosophy and Phenomenological Research*, **55**, 1, pp. 109–131.
Thomas, A. 2006, *Value and Context*, Oxford: Oxford University Press.
Williams, B. 1981, "Internal and External Reasons", reprinted in: B. Williams, *Moral Luck*, Cambridge: Cambridge University Press.
Williams, B. 1989, "Internal Reasons and the Obscurity of Blame", in: B. Williams, *Making Sense of Humanity*, Cambridge: Cambridge University Press.
Williams, B. 1995. 'Replies', in J. Altham and R. Harrison, eds, *World, Mind and Ethics*. Cambridge: Cambridge University Press.
Wong, D. 2006, "Moral Reasons: Internal and External", *Philosophy and Phenomenological Research*, **72**, 3, pp. 536–558.