On the disconfirmation of practical judgements*

Sander Voerman

2009-12-12

When we have to make practical judgements—judgements about how to live our personal lives, what sort of careers to pursue, what moral and political views to adopt and how to act upon them, and so forth—we are concerned to *get those judgements right* (Smith, 1994, p. 5; Frankfurt, 2006, p. 2, 27). In this essay, I discuss a question which is closely related to this concern: the question of how we can discover that we got our practical judgements *wrong*. According to the "Affective Response View" that I will be proposing, we disconfirm our practical judgements on the basis of *affective experiences* that we *did not expect* ourselves to have. This view avoids certain problems of the "Principles of Reason View," the familiar view that we disconfirm practical judgements by showing that they violate *a priori* principles of reason. My proposal implies a relativistic form of realism about normative reasons for action: practical judgements are true or false in virtue of empirical facts about the agent's "normative will," a pattern of dispositions that determines her motivations under ideal conditions of rational agency. I will argue that this view is attractive for several reasons, one of them being that it allows us to explain how the disconfirmation of a practical judgement *motivates* a self-governing agent to change her behaviour accordingly.

1 Two questions about practical belief

In order to account for the concern to get our practical judgements right, we might want to subscribe to *cognitivism*, the view that practical judgements express or establish *beliefs*: if A judges that he should ϕ, then A believes that he should ϕ. Let us call such beliefs "practical beliefs." Most cognitivists are *realists*: they subscribe to the stronger claim that there are facts which make certain practical beliefs *true*. Cognitivism offers an explication of the intuition that we are concerned to get our practical judgements right: we

*Thanks to Bert van Roermund, Herman de Regt, Marc Slors, Derek Strijbos, Katrien Schaubroeck and an anonymous referee for *Logique et Analyse* for their helpful comments. Work on this essay was sponsored by the Dutch Organization for Scientific Research (NWO).

are concerned to adopt true practical beliefs. Realism allows us to be non-sceptical about that concern: we *can* get our practical judgements right because there are facts about what we should do.[1]

A question that many realists have tried to answer is: what sort of facts might that be? What sort of facts make it true that I should keep my promise, or that John should help the woman that just fell from the stairs? Let us call this the "fact question."

There is a second question that realists should answer. Note that the concern to get practical judgements right makes it reasonable to assume that we can get them *wrong* (otherwise there would be no need to be so concerned). Consider the Montgomery bus drivers who, before the Boycott of 1955, forced black passengers to give up their seats for white passengers. Suppose that one bus driver, who used to judge that it was right for him to enforce this policy, heavily revised his views later on in his life, to the point where he would forcefully advocate racial equality and the abolishment of any such policies. It seems plausible that such a revision involves more than a change in preference. From a realist point of view, we want to be able to say that the bus driver *discovered* that his practical beliefs were *false*.

But how did he do that? That is the second question that realists must answer. How do we *disconfirm* practical beliefs? What sort of consideration makes it rational for an agent A to reject her belief that she should ϕ? I shall call this the "disconfirmation question."

The purpose of this essay is to discuss how we might answer the fact question and the disconfirmation question if we accept a certain version of *internalism*, the view that practical judgements have motivational implications. The attempt to give internalistic answers to these questions leads to what I shall call the "motivation problem," a central problem in meta-ethics about the relation between normativity and motivation. This problem is usually discussed in the context of the fact question, which may appear to be the most fundamental of the two questions. I shall briefly rehearse this discussion in the next section, and formulate what I take to be the most promising internalistic answers to the fact question in the literature. However, in my own view it is much more fruitful to think of the *disconfirmation question* as the primary question, and to approach the fact question as derivative. I shall therefore reformulate the motivation problem in the context of the disconfirmation question (section 3). I will then propose my

[1]There are, of course, several alternatives. Inferentialists allow that some practical beliefs are true, but deny they are made true by facts. Error theorists claim that all practical beliefs are false. And non-cognitivists deny that practical judgements express beliefs in the first place. Although realism seems to give the most straightforward justification of our concern to get practical judgements right, I shall not be arguing that anti-realists cannot accommodate this concern. Instead, my purpose in this paper is to proceed from the assumption that realism is true and to focus on certain problems that arise from this assumption.

theory of practical disconfirmation, the "Affective Response View" (sections 4 and 6), explore its implications for the fact question (section 5), and explain how this proposal solves the problem of motivation (section 7).

2 Can we give an internalist answer to the fact question?

The version of internalism that I wish to discuss is as follows: if A judges that she should ϕ, then it follows with conceptual necessity that A has what I shall call a "self-adopted reason" to ϕ: she is either sufficiently motivated to ϕ (she has a "motivating reason" to ϕ), or insofar she lacks that motivation, this is due to an impairment in her self-government, such as a compulsive disorder or weakness of will (Blackburn, 1984; Harman, 1978/2000b, p. 30; Smith, 1994, pp. 60–63). Internalism accommodates the intuition that when we make a practical judgement, we exercise a kind of *authority* over ourselves. This intuition explains the normative character, the "demandingness" of such judgements: for what could this normative character possibly consist in, we may ask, if the person who *makes* the judgement would herself not feel required to live up to it?[2]

Note that the concept of "being *sufficiently motivated* to ϕ" allows that one also has a desire *not* to ϕ, or that one experiences other negative feelings about ϕ-ing, as long as the resultant force, so to speak, of the totality of one's affective attitudes towards ϕ, is positive. Let us call such a positive resultant attitude a "resultant desire" to ϕ, or a desire to ϕ in "the resultant sense." Thus, the resultant desire may incorporate a multiplicity of desires, but also sensation responses such as pain and pleasure, and emotions such as regret or jealousy: all states that contribute motivating impetus. Where ϕ is a concrete action that the agent is capable of performing, having a resultant desire to ϕ means that she *will* ϕ. However, we may also want to adopt internalism with respect to practical judgements about political ideals or states of affairs that the agent does not have the power to immediately bring about, or even to influence at all. In that case, ϕ may mean something like "supporting P," "contributing to P," or, simply, "approving of P," and we can formulate internalism as follows: if A judges in approval of P, then it follows with conceptual necessity that either A has a resultant desire *that* P, or A is impaired in her self-government.

If cognitivism is true, then internalism implies that practical judgements have a dual nature: they express or establish both practical beliefs and self-adopted reasons for action. The internalistic cognitivist may want to identify the two and simply conclude that on his combined view, practical beliefs *are* self-adopted reasons for action. The implication for realism is

[2]But see Brink (1986) and Watson (1987/2004, pp. 168–169) for possible answers to this rhetorical question.

as follows. Suppose that some fact makes it true that A should ϕ. Then A could not come to believe this truth without adopting it as a reason for himself to ϕ. "Normative truths," as Harry Frankfurt puts it, "require that we submit to them" (2006, p. 34). Following Michael Smith, let us call such truths "normative reasons for action" (Smith, 1994, p. 94). Internalistic realism, then, is the view that there are normative reasons for action.

This view has troubled philosophers who agree with a principle from David Hume, that merely believing something does not by itself generate or require a motivation to do anything (1886/1964). If some fact makes it true that I have a normative reason to ϕ, then such a fact may seem to violate this principle, since my knowledge of such a fact would have to motivate me to ϕ if I were fully self-governing. The internalistic realist must explain how this is possible: how could any sort of fact, upon being known by a self-governing agent, determine how that agent is motivated? That, in a nutshell, is the motivation problem.

In order to solve this problem, various authors have tried to defend a "dispositional" or "response-dependence" theory of practical normativity. According to such a theory, we have normative reasons for action in virtue of facts about what our motivations and affective responses would be under certain ideal conditions of rational agency (Firth, 1952; Williams, 1980/1981; Smith, 1989, 1994, 2002/2004a; Lewis, 1989; Johnston, 1989; Jackson & Pettit, 1995). These conditions typically include self-government, flawless reasoning, and access to all the relevant information. The general idea is simple: we remove the mystery about why we would be motivated, under ideal conditions, in accordance with our normative reasons, by *analyzing* normative reasons in terms of how we would be motivated under those conditions.

Smith distinguishes a "non-relativistic" from a "relativistic" version of dispositionalism (1995/2004b, pp. 25–34; 2000/2004c, pp. 204–206). The non-relativistic version states that an agent has a normative reason to ϕ under circumstances C if and only if *all agents* would, under ideal conditions of rational agency, and when familiar with circumstances C, desire in the resultant sense that *any agent* would ϕ under those circumstances. Note that the circumstances C may include the tastes and preferences of the agent: thus, all agents might under ideal conditions desire that those who prefer tennis play tennis whereas those who prefer basketball play basketball. Note also that the circumstances C may be inconsistent with the ideal conditions of rational agency. For example, C might include the fact that the agent in question has difficulty to control his anger. Presumably, under the ideal conditions of rational agency he would no longer have this difficulty. It follows that what an actual agent has normative reason to do is not necessarily the same as what he himself under ideal conditions of rational agency—his "ideal self" as Smith has called it—would do. Rather, it is what his ideal self would *advise* his less than ideal, actual self to do

(Smith, 1995/2004b, pp. 18–20). Finally, note that the phrase "all agents" in the formulation above refers to *all conceptually possible agents*. It follows that if we have normative reasons for action, then there are resultant desires that all conceptually possible agents would share under the ideal conditions of rational agency, which means that the content of those desires would be determined *a priori* by those conditions alone. Hence, according to the non-relativistic dispositionalist, the facts that make it true that we have normative reasons for action are a kind of "conceptual facts" about rationality (this is roughly the view of Smith; for related views, see Firth, 1952; Korsgaard, 1986; Jackson & Pettit, 1995).

In contrast, according to the relativistic version of dispositionalism, all resultant desires of the ideal self of an agent A would be *functions* of A's actual contingent motivational characteristics. On this view, every agent has her own "normative will," as I shall call it: her own source of normative reasons for action, which depends on her actual characteristics even though it would only fully manifest itself under the ideal conditions of rational agency. Let us apply Smith's advice-interpretation to this view as well: A has a normative reason to ϕ under circumstances C if and only if the ideal self of A would have the resultant desire that his actual self would ϕ under C, but the ideal selves of other agents might desire differently. This is still a form of realism in the sense defined above: for any particular agent, there are facts about what that agent has normative reason to do. However, these are not conceptual facts about rationality, but *empirical* facts about *that particular agent*. Therefore, the relativist denies that certain actions under certain circumstances have the objective property of being "right" or "good" in the sense that *every* agent should approve of those actions, when familiar with those circumstances, regardless of his own actual attitudes. In that sense, the relativist is an anti-realist (this is basically my own view; similar views can be found in Williams, 1980/1981; Harman, 1985/2000a; and Frankfurt, 2006).

An elaborate discussion of the debate between relativism and non-relativism is beyond the scope of this paper. However, as we shall see below, my proposal on the issue of disconfirmation will have relativistic implications. Let us now turn to the disconfirmation question.

3 Can we give an internalist answer to the disconfirmation question?

In the light of the disconfirmation question, we can reformulate the motivation problem as follows. If internalism is correct, then it is conceptually necessary that a *change* in our practical views implies a *corresponding* change in our self-adopted reasons for action. For the realist, this means that if an agent A realizes that X disconfirms her belief that she has a normative

reason to ϕ, and requires her to adopt the belief that she has a normative reason to ψ instead, then in the light of X, she would not only be irrational if she failed to change her beliefs accordingly, but she would also be lacking in self-government if she would not lose her resultant desire to ϕ and gain a resultant desire to ψ. The question is, what sort of X might have this dual impact on her attitudes?

Cases of instrumental reasoning are easy: suppose that X is evidence that ψ, rather than ϕ, would allow A to accomplish ω. If A had a derived desire to ϕ in order to fulfill her intrinsic desire to ω, she may be expected, upon learning of X, to lose her desire to ϕ and acquire a desire to ψ instead. The problem is how to answer the disconfirmation question in non-instrumental cases. If we assume that intrinsic desires are entirely non-cognitive attitudes, which are not subject to matters of belief, then how could there be any X such that X would both disconfirm a belief and diminish an intrinsic desire of a self-governing agent?

Even though this problem is in some sense the same problem that we discussed in the context of the fact question—because it arises out of the same tension between realism and internalism—it is not obvious from the dispositionalist answer to the fact question how we must solve the problem in its reformulated form. Dispositionalism makes a claim about agents under *ideal conditions* of rational agency, conditions that are never actually fulfilled, which gives the dispositionalist a lot of room for speculation about what might be true under those conditions. However, in order to give an internalist answer to the disconfirmation question, we must explain how the progress of our understanding of our normative reasons for action could be connected to changes in our motivations *as a matter of actual fact.*

This challenge has implications for the dispute between relativistic and non-relativistic dispostionalism. In particular, only solutions that satisfy two additional criteria are consistent with non-relativistic dispositionalism. These criteria are, first, that valid practical disconfirmations give rise to motivational *convergence* for all conceptually possible agents, and second, that valid non-instrumental practical disconfirmations are justified on *a priori* grounds.

Let me explain. Recall that according to the non-relativistic dispositionalist, if there is a normative reason for some agent to ϕ under circumstances C, then for every conceptually possible agent T it must be true that the ideal self of T would desire that agents ϕ under C. This implies that if we have normative reasons for action, then there would have to be a class V of propositions of the form "agents ϕ under C" such that the ideal selves of all conceptually possible agents desire all propositions in V. However, the non-ideal selves of all conceptually possible agents vary in their desired propositions in all conceptually possible respects. Non-relativistic dispositionalism implies that the desired propositions of these agents will *converge* onto classes containing V, once these agents start approaching their ideal

selves by disconfirming their false practical beliefs. The non-relativistic dispositionalist must therefore explain what sort of X could both disconfirm the beliefs of different conceptually possible agents and at the same time make their motivations converge onto the same desires.

Furthermore, the role that empirical facts may play in practical disconfirmation is limited for the non-relativistic dispositionalist to purely instrumental concerns. Suppose that A believes that he should ϕ under the present circumstances, because he believes (a) that he should achieve ω under circumstances C, (b) that the present circumstances are of type C, and (c) that ϕ-ing would be a way to achieve ω. Then his beliefs (b) and (c) may be disconfirmed by *a posteriori* knowledge—either of the empirical fact that the present circumstances are not of type C, or of the empirical fact that ϕ is not a way to achieve ω. However, suppose that A would rule out all errors of these kinds. Then it seems that A may still be mistaken in his self-adopted *ends*—his practical beliefs of type (a) that are not instrumental derivations from other such practical beliefs. If non-relativistic dispositionalism is true, then false practical beliefs of this kind cannot be disconfirmed by *a posteriori* knowledge of the actual world, because agents in remote conceptually possible worlds, where the empirical facts are entirely different, would have to be able to disconfirm those practical beliefs as well. Therefore, non-relativistic dispositionalism implies that non-instrumental disconfirmation of practical beliefs must be *a priori*: it must be determined solely by what it would mean for any agent to desire something under ideal conditions of rational agency.

What sort of account would satisfy these two criteria? Smith argues that we revise our practical beliefs on the basis of rational reflection in the light of certain *principles of reason* (2007, pp. 136–140). These might include principles of universalization, for example, or other principles of coherence that go beyond the meager means-end coherence principle of instrumental disconfirmation. Suppose one could make it plausible that some principle of reason is constitutive of the ideal conditions of rational agency, in the sense that an agent cannot violate the principle under those conditions. In that case, showing that a practical belief violates the principle would be an *a priori* ground for disconfirming the practical belief. Let us call the view that such principles underlie our practical disconfirmations the "Principles of Reason View." This view has a certain intuitive appeal: many moral philosophers, after all, have tried to formulate such principles (Kant's categorical imperative being the most notable example). Nevertheless, even if there are such principles, then it seems doubtful, due to their formal nature, that they can account for the disconfirmation of all our mistaken ends. What seems truly *stunning*, however, is that such formal principles would be able to settle conflicts of desire *between* agents and lead all agents toward the same intrinsic desires no matter what their initial intrinsic desires were. Therefore, various authors—including Smith himself—have expressed

scepticism about the idea that the Principles of Reason View could deliver the convergence that the non-relativistic dispositionalist is committed to (Sobel, 1999; Enoch, 2007, p. 106; Smith, 2006, pp. 77, 102; 2007, pp. 136–137).

I shall not argue in further detail that these problems for the Principles of Reason View cannot be solved. My purpose in this essay is to propose an alternative view of disconfirmation, a view from which it follows that even non-instrumental practical disconfirmation is *a posteriori*, and which therefore leads to a relativistic version of dispositionalism. Of course, this may be a reason for non-relativists to reject it outright, but as I hope to show, my proposal may be plausible for independent reasons. Furthermore, as we shall see, the implied version of relativism has a number of attractive features.

4 The Affective Response View

According to the view that I want to propose, our practical beliefs can be disconfirmed by our own *affective responses* to our self-governed actions, or to the intended consequences of those actions, *insofar as we did not expect ourselves to experience those responses*. If a thief judges that he has a good reason to steal from someone, and does not expect himself to feel very guilty about it, then he may come to doubt his initial judgement if after the theft he gets overwhelmed by feelings of guilt. On my proposal, this is because such feelings have the power to disconfirm practical beliefs. I call this the "Affective Response View."

First, let me explicate the notion of an "affective response." By this I mean any affective attitude, experience or sensation that can be understood as a response to an event that preceded it. This might be any sort of event, but our discussion concerns affective responses to actions or the consequences of actions. Whether an affective experience should be understood as a response to a certain event may be a subject of interpretation. If the affective experience is an intentional attitude of remorse about having murdered someone then the experience is clearly a response to the murder, but if the affective experience is a general feeling of joy without any specific content, then it may not be clear whether or not this is a response to a certain previous act or event (I will say more about the interpretation of responses in section 6).

Since affective responses are backward-looking attitudes, as it were, they may be thought of as a kind of counterparts to desires, which are typically forward-looking. The feeling of satisfaction as a result of an action is such a counterpart to the feeling of desire that motivated the action. However, sometimes forward-looking desires may also be understood as affective responses themselves. For example, suppose one decides to become a vegetar-

ian. If, subsequently, one's desire for meat increases, this may be interpreted as a response to the (consequences of the) decision. Furthermore, every affective response to an event may be understood as a desire in a very loose sense—as the desire that P, where P is the proposition that the event occurred (in the case of a positive response) or the proposition that the event did not occur (in the case of a negative response). Finally, every affective response contributes to the *resultant desire* of the agent at the time of the response in the sense defined in section 2 above. For example, suppose that I have eaten seven slices of pizza and I am wondering whether or not to eat the eighth slice. My affective attitudes might be mixed. On the one hand, I desire to eat it because I want to taste some more. On the other hand, the way my stomach feels tells me that I have already eaten too much. This affective response may outweigh the desire to eat the last slice and make a decisive contribution to my resultant desire not to finish the pizza. Hence, affective responses can be efficacious motivational states.

Let us now turn to the role that affective responses play in practical disconfirmation. Common examples of affective responses that sometimes make us rethink our prior judgements are feelings of regret, remorse, guilt, shame, embarrassment, jealousy and boredom. Nevertheless, disconfirmation is not intrinsic to these responses. Rather, whether an affective response disconfirms a prior judgement depends on how that response is related to other affective states. The general idea behind the Affective Response View is that self-governing agents will expect a kind of "match" between the affective states that motivate their actions on the one hand, and their overall affective responses to the consequences of those actions on the other hand. If John desires to see Rome, and judges that he should spend his money on a holiday to Italy's remarkable capital, then he will expect his visit to Rome to be a pleasurable and rewarding experience. If the holiday would fail to meet these expectations, then John might start to think that his money would have been better spent differently.

However, we will rarely expect our responses to completely match the desires that we decided to act upon. If Carol deliberates about whether or not to quit her job and accept another one, then she will probably both have desires in favour of quitting and desires in favour of staying. Should she decide to make the change, then in the light of her multitude of desires, she may expect both positive and negative responses. In the short term, she might even expect the negative responses to be stronger because of the stress and the various difficulties of adjustment. Nevertheless, it seems plausible that if she decides to quit, her expectation will be for her overall response to be more positive *in the long run* than if she would have stayed. Should her actual responses, after a while, give her reason to believe that she would have been happier if she had kept her old job, then her responses may disconfirm her prior judgement.

This does not mean that the Affective Response View commits us to a he-

donistic egoism about maximizing one's own happiness—at least not under any shallow interpretation of the terms "hedonism", "egoism" or "happiness." Suppose that Jack is in a hurry, and decides not to help an injured person on the street. If Jack would feel ashamed of himself afterwards, and if he were to conclude that he should have helped the injured person, then a defender of the Affective Response View might argue that the feeling of shame disconfirmed Jack's prior practical belief and made him adopt the practical belief that he had a normative reason to help the injured person. But that does not mean that Jack merely had a normative reason to do so in order to prevent himself from feeling bad about himself, which would have been a purely instrumental consideration. Rather, it means that his feeling of shame informed him of the fact that helping the injured person was more important to him than he initially thought.

5 The relativistic implication

Although the Affective Responsive View does not commit us to shallow egoism or hedonism, it does imply *relativism*. The practical beliefs of agent A about what she has normative reason to do are subject to disconfirmation by *her* affective responses, which may tell her something about what is important to *her*. Perhaps the non-relativist might want to object that affective responses could be intuitions about *a priori* principles of reason, which would carry us back to the Principles of Reason View. But it is not clear how this suggestion would make the problems for the Principles of Reason View any easier. Our affective responses are a result of contingent psychological mechanisms, and without a proper explanation of why their content would involve *a priori* truths about reasons for action, we have no reason to think that they constitute anything other than a matter of empirical fact. *Prima facie*, for any agent A, the content of A's affective responses seem to give empirical information about A, rather than *a priori* information about all conceptually possible agents.

Thus, suppose that Sharon is a vegetarian. She becomes friends with Marc and David, who are both used to eating meat and never felt bad about it. However, once Marc gets to know Sharon better, and starts to consider things from her perspective, he discovers that he begins to experience negative feelings about eating meat. He starts to feel guilty about the idea that animals were killed in order for him to enjoy a particular eating habit, even though that habit is not necessary in order for him to live a healthy life. On the basis of these feelings, Marc starts to disapprove of the killing of animals by humans for food, thereby disconfirming his prior belief that it was okay to do so. Should it be the case, in virtue of *empirical facts about Marc*, that this is also how Marc would feel under ideal conditions of rational agency, then it does not follow that *any conceptually possible agent*

would have to feel the same under those conditions. Thus, suppose that David does not develop negative feelings about killing animals for food at all, not even after extensive discussion with Marc and Sharon. Under ideal conditions of rational agency, David may still have a resultant desire to eat meat, even though Marc and Sharon may have the resultant desire under those conditions that nobody would eat meat.

However, it may well be an empirical fact about human psychology *in general* that there are certain actions that all of us do have normative reason to disapprove of. For example, it may well be an empirical fact that every human being would under ideal conditions of rational agency have the resultant desire that no sentient being ever be tortured. Therefore, the Affective Response View does not prevent us from arguing, say, that the Nazis got their practical judgements wrong. Psychologically, the Nazis had so much in common with us that it seems plausible that under different circumstances, they would have had the same feelings of horror about the Holocaust that we do. We may ascribe the fact that they did not actually feel this way to a type of upbringing and training that, effectively, removed them further away from the ideal conditions of rational agency, and thereby made it impossible for them to fully understand their own affective dispositions. In other words: every SS officer who believed that he had a normative reason to torture and murder his victims may have gotten *himself* wrong.[3]

What the Affective Response View does rule out is that every conceptually possible agent would get it wrong when judging in approval of torture and genocide. Thus, suppose that aliens from outer space would invade our planet and start torturing and exterminating us. If, as a matter of empirical fact, these aliens would have no disposition whatsoever to sympathize with us, then it will be impossible, on the Affective Response View, to disconfirm their practical beliefs. This is where the Affective Response View differs from the Principles of Reason View. But as we have seen in section 3, the implication that all conceptually possible agents would have to be able to disconfirm their practical judgements in such a way as to end up having the same practical views is actually a *problem* for the Principles of Reason View. Therefore, the fact that the Affective Response View does not have this implication seems to me to be a *feature* rather than a bug. What this means is that if the Nazis got their practical judgements wrong, they got it wrong precisely because they were *not* alien monsters: they got it wrong *as human beings*.

Allow me to introduce some additional terminology at this point. According to the type of relativism that we have been discussing, every agent

[3]The idea that his training prevented the SS officer from understanding his ideal self raises the question of whether it would have been possible to make that training undone. If not, then we may wonder what sort of counterfactual life histories we should consider in order to construe his ideal self. At what point would it become the ideal self of a *different* person? This problem requires a separate discussion elsewhere.

has his own source of normative reasons for action, which consists in certain empirical facts about his psychology. Let us call this source the "normative will": let us say that A wants to ϕ under circumstances C in the "normative sense," or that ϕ-ing under C is "part of the normative will" of A, if and only if under the ideal conditions of rational agency, A would desire in the resultant sense that under the circumstances C, he would ϕ. One may think of the normative will as a complex of the agent's deepest attitudes of caring and love, which establish what is most important to him (Frankfurt, 2004, 2006). However, note that these attitudes may be phenomenologically *opaque*. The normative mode of wanting is a mode of wanting that we may ourselves be ignorant of: at the time, Jack did not know that he wanted, in the normative sense, to help the injured person, and SS officers did not know that in the normative sense, they did not want to torture and kill their victims.

6 Will interpretation

I have claimed that a self-governing agent will expect the intended consequences of her actions to generate the most positive overall affective response, in the long run, compared to the alternatives that she might have chosen. The underlying intuition is that in the long run, our overall responses to our actions tell us something about what we want in the normative sense: that they will approach the resultant desires of our ideal selves, so to speak. However, the notions of "overall response" and "in the long run" are of course totally vague and abstract. In practice, our responses change from moment to moment and from situation to situation, and it is often hard to determine which of our affective experiences are responses to which consequences of our actions. Therefore, whether an affective experience disconfirms a practical belief is always a matter of *interpretation*: we must judge what the experience *means* to us.

Suppose, for example, that a student feels an unexpected embarrassment after having asked a question during a course meeting (perhaps it turned out that it was not a very intelligent question). Does that mean that he shouldn't have asked it? Perhaps it does, but perhaps it doesn't. The student might also conclude that this merely reveals that his questions can be as unintelligent as those of anybody else, and that perhaps he may be more easily embarrassed about this than he thought he would be. But since he won't get any smarter by not asking such questions, he might reason for himself, it was still a good idea to ask the question anyway. In other words: the negative response of embarrassment about his action does not *intrinsically* have higher normative authority than the positive affect of curiosity that initially motivated the action.

In fact, the judgement that an agent must make in order to determine

whether a response disconfirms an action is of exactly the same kind as the judgement that he had to make before the action in order to determine whether he wanted to act upon the desire that motivated the action: it is just another practical judgement, a judgement about whether the affective experience is an expression of his normative will. Our concern to get our practical judgements right brings with it a concern to know whether our affective responses are appropriate or not. The point of the Affective Response View, however, is that it is only on the basis of *other* affective responses that an agent could be justified in judging that his current affective response is *in*appropriate.[4]

Let me illustrate. Suppose that the student does conclude that he shouldn't have asked the question. The next time that his curiosity boils up, he remembers the unpleasant embarrassment, which is itself an unpleasant experience that counteracts his motivation to ask another question. Suppose that he decides not to ask the question this time, and that he is self-governing—i.e., the unpleasantness of his memory of the previous time is stronger than his desire to ask another question. By not asking the question, he might save himself another embarrassment, but when the course is finished, his curiosity is unsatisfied, which makes him feel frustrated. Perhaps, again, more than he would have expected. And again, this is an experience that requires interpretation. What does it mean? Perhaps it means that he should visit the *Wikipedia* and try to find the answer to his question for himself. But it might also mean that he does not want his fear of embarrassment to prevent himself from asking what he really wants to know, and that he should have asked the question after all. In that case, the judgement that the embarrassment was a disconfirmation would itself be disconfirmed, and the initial judgement which led him to ask the first question would be *confirmed*. This shows that deliberation is an ongoing process of what I shall call "will interpretation": the interpretation of our affective experiences in order to determine which of them express our normative will.

Let me make a number of brief remarks about this notion of will interpretation. First of all, note that we are now dealing with two directions of explanation. The moral philosopher who wants to answer the *fact question* is interested in knowing whether we can explain the existence of normative

[4]Some people might also believe that *most* of their affective responses are inappropriate, that their *overall* affective response is inappropriate, or perhaps even that their overall response would still be inappropriate under ideal conditions of rational agency. An unmarried man might have an overall affective response in support of his promiscuous life, for example, and still believe that he should not be living such a life, because his religion teaches that sex is only allowed within marriage. Such an agent would have to reject the Affective Response View, and perhaps even the more general thesis of internalism. But that does not mean that internalism or the Affective Response View are false. It only means that if these views are true, then given that his overall response is in favour of his lifestyle, his practical judgement against it must be, to borrow a phrase from Williams, "false, incoherent, or really something else misleadingly expressed" (1980/1981, p. 111).

reasons for action in terms of facts about the affective attitudes of agents. But the idea of will interpretation is that *as deliberating agents*, we often reason in the opposite direction: we want to know whether a certain reason for action would explain the affective attitudes we are experiencing. If I have a normative reason to ϕ, after all, then insofar I am rational, well-informed and in control of my own agency, I should expect to find a certain pattern in my actual experiences in support of ϕ. Thus, the hypothesis that I have such a reason may be explanatory relevant, in a structural sense, to my actual motivation to ϕ. Metaphysically, then, the nature of normative reasons for action will be something like *patterns* or *structures* of affective dispositions. Further inquiry into the metaphysics of the normative will is beyond the scope of the present paper, however.

A second point that I want to highlight is that the analysis of deliberation as will interpretation allows us to understand deliberation as a *social practice*, even though the core of the analysis is individualistic. For one thing, different people often display similar responses in similar situations as a result of underlying psychological structures that we all have in common due to our shared environment and biological ancestry. Therefore, we can learn from each other's mistakes, and we can search for "human values"—things that all human beings would want under the ideal conditions of rational agency as a result of the empirical facts about human psychology. This allows us to argue that the Nazis got their practical judgements wrong, for example, as I have outlined in the previous section. Furthermore, once you come to believe that you have even more in common with a specific group of people, or with a particular person, then it becomes even more plausible to expect similar responses in the situations which pertain to those commonalities.

However, we may also improve our interpretation of what we really want on the basis of discussion with those who want something different. Such discussion often forces us to articulate more precisely the reasons that we have for our practical beliefs, which may lead us to revise those beliefs in subtle ways and to make them more sophisticated and precise. Furthermore, there may be cases where someone close to me understands what I really want before I understand it myself, even though it need not be something that *she* really wants.

In fact, we often do not take alternatives seriously until they are being demonstrated or suggested to us by certain individuals, groups, or media. In the absence of any social pressure to change their views, people generally stick with their initial gut feelings, and self-confirmation bias is everywhere in our psychology (Haidt, 2001). That is why we rarely disconfirm our beliefs about what we care about most. The problem is not just that we protect our self-image by ignoring evidence, or by giving heavily biased interpretations of unexpected affective responses. The problem is also that we rarely experience unexpected responses concerning our most cherished practical

beliefs in the first place, because our responses do not arise independently of those beliefs.[5] Thus, it is possible that an agent experiences no affective responses against ϕ-ing, and that there is no doubt in his mind that he has a normative reason to ϕ ("he knows what he wants"), while his normative will is actually opposed to ϕ-ing, in virtue of the fact that he would eventually experience massively adverse responses with regard to ϕ-ing, once he would start taking the possible reasons not to ϕ more seriously.

Therefore, will interpretation may benefit from attempts to break out of dogmatic self-assumptions, and social influence can be a way of making people consider new alternatives. Of course, in reality, social practice often only makes things worse, because people will prevent each other from starting to doubt the views that constitute the identity of their group. Nevertheless, many of our most fundamental changes in our practical beliefs have occurred in the context of social developments. The case of the Montgomery Bus Boycott and the rise of the civil rights movement would be an example of such a development. Thus, we might argue that as a result of this development, people like our fictitious bus driver did not just start taking a different point of view seriously, but they also started experiencing different affective responses to established practices, including their own actions, which eventually led to the disconfirmation of their segregationist practical beliefs.

A related point about will interpretation is that it never reaches *final* results. Practical beliefs, on this view, are always *hypothetical*: they are forever subject to revision in the light of new experience. This does not rule out that an agent may have good reason to be "fully resolved" in some of his practical judgements, as Harry Frankfurt has put it, in the sense that the agent may have the "belief that no further accurate inquiry would require him to change his mind" (1987/1988, p. 169). Sometimes we *do* know what we want. For example, nowadays we may well believe that no inquiry will ever disconfirm our practical belief that people should be treated equally regardless of the colour of their skin or their sexual orientation, say. Nevertheless, in the light of the theory of will interpretation, it would probably be wise for most of us to keep an open mind and to take alternatives to *most* of our present practical beliefs seriously. Furthermore, the hypothetical nature of will interpretation *does* rule out Frankfurt's notion that any particular affective experience could reveal a "volitional necessity," a directly experienced constraint, imposed by the normative will of a person, on what he can and cannot bring himself to do (2004, pp. 46–49; 2006, pp. 33–34). Instead, on my account, if we cannot bring ourselves to do what we had judged that we should do, then it is always a matter of interpretation whether we are

[5] An analogy may be drawn with issues concerning the theory-ladenness of observation in the philosophy of science. This analogy is pursued, to some extent, in Churchland (1995, ch. 6, esp. p. 146–147). See also his (1989, pp. 188–196) on the theory-ladenness of observation.

experiencing disconfirmation or merely weakness of will.

Another point to note about the concept of will interpretation is that it is compatible with different methods and approaches to deliberation, and that it also allows us to *combine* those approaches, which I think is a very attractive feature. Thus, it might be that in certain areas of ethics and political theory, it will be very useful to try to formulate individual or shared practical beliefs using *principles*, such as principles of universalizability. On other moral issues, however, we might have better luck if we try to describe *values*, or if we reflect on how we might embody certain *virtues*. And in certain domains of our personal lives, a *narrative* approach would perhaps be most suitable: we deliberate by articulating the story that we would want to tell about ourselves. From the point of view of will interpretation, these are merely different *interpretative strategies*, and every strategy is valid as long as it yields disconfirmable expectations about our future affective responses.

7 The motivation problem solved

Let us now return to the 'disconfirmation version' of the motivation problem. According to the Affective Response View, our practical beliefs may be disconfirmed by unexpected affective responses. So why should that result in a non-instrumental change in our motivations? The answer is that the unexpected affective responses *are* the changes in our motivations. After all, affective responses are themselves motivational experiences, which influence our future behaviour. If they come unexpected, then it is likely that we were not used to having them, which may signify a change, and alter the balance of affective attitudes in such a way as to lead to a new resultant desire. If we judge that they nonetheless express what we really want, then the motivational change and the belief change may correspond to each other in such a way that we maintain the same level of self-government.

Recall the example of the student who got embarrassed after having asked the unintelligent question. Suppose that the student judged that he shouldn't have asked the question, and decides that he won't ask such a question again. Then the embarrassment, through his unpleasant memory thereof, will have changed his motivational disposition, and—if it is stronger than the curiosity—will prevent him from asking another question.

Note that this solution is very much in the spirit of the dispositional theory of practical normativity. We remove the mystery about why the disconfirmation would imply a motivational change, under conditions of sustained self-government, by claiming that under those conditions, the disconfirmation simply *is* the motivational change.

8 Conclusion

My aim in this paper was to propose the Affective Response View as an answer to the disconfirmation question, and to situate this view within the debate on the motivation problem, which is often focused primarily on the fact question. Unavoidably, the details of the proposal have remained somewhat sketchy. Nevertheless, there are a number of things to be said in favour of the proposed view.

First, as I have just argued in the previous section, it offers a solution to the motivation problem in the context of the disconfirmation question.

Second, I think the view appeals to a common sense intuition about disconfirmation in practice. For it seems common sense that emotional responses like shame and regret play a role in our practical re-evaluations. But it has always been a challenge for *a priori* accounts of disconfirmation to explain an intuitive connection between reflection on *a priori* principles on the one hand, and contingent empirical facts about our emotional dispositions on the other hand. Instead, the Affective Response View takes the intuitive role of affective responses in practical disconfirmation at face value.

Furthermore, I have argued that the Affective Response View allows us to account for intersubjective aspects of moral discourse and debate, even though its implication for the fact question is relativistic dispositionalism. Thus, the proposal is meant to contribute to the development of a sophisticated form of relativism, which is neither egoistic in its account of the will, nor solipsistic in its account of practical reason.

Tilburg University
s.a.voerman@uvt.nl

References

Blackburn, S. (1984). *Spreading the word: Groundings in the philosophy of language*. Oxford: Clarendon Press.

Brink, D. O. (1986). Externalist moral realism. *Southern Journal of Philosophy*, *24*(Suppl.), 23–42.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: The MIT Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA: The MIT Press.

Enoch, D. (2007). Rationality, coherence, convergence: A critical comment on Michael Smith's *Ethics and the A Priori*. *Philosophical Books*, *48*(2), 99–108.

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, *12*(3), 317–345.

Frankfurt, H. G. (1988). Identification and wholeheartedness. In *The importance of what we care about: Philosophical essays* (pp. 159–176). New York: Cambridge University Press. (Reprinted from *Responsibility, character and the emotions: New essays in moral psychology*, by F. D. Schoeman, Ed., 1987, New York: Cambridge University Press)

Frankfurt, H. G. (2004). *The reasons of love.* Princeton: Princeton University Press.

Frankfurt, H. G. (2006). *Taking ourselves seriously & Getting it right* (D. Satz, Ed.). Stanford: Stanford University Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgement. *Psychological Review*, *108*(4), 814–834.

Harman, G. (2000a). Is there a single true morality? In *Eplaining value and other essays in moral philosophy* (pp. 77–99). New York: Oxford University Press. (Reprinted from *Morality, reason and truth*, pp. 27–48, by D. Copp & D. Zimmerman, Eds., 1985, Totowa, NJ: Rowan & Littlefield)

Harman, G. (2000b). What is moral relativism? In *Eplaining value and other essays in moral philosophy* (pp. 20–38). New York: Oxford University Press. (Reprinted from *Values and morals*, pp. 143–161, by A. I. Goldman & J. Kim, Eds., 1978, Dordrecht: Reidel)

Hume, D. (1964). A treatise of human nature. In T. H. Green & T. H. Grose (Eds.), *David Hume: The philosophical works.* Aalen: Scientia Verlag. (Original work published 1886)

Jackson, F. C., & Pettit, P. (1995). Moral functionalism and moral motivation. *The Philosophical Quarterly*, *45*(178), 20–40.

Johnston, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 139–174.

Korsgaard, C. M. (1986). Skepticism about practical reason. *Journal of Philosophy*, *83*, 5–25.

Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 113–137.

Smith, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 89–111.

Smith, M. (1994). *The moral problem.* Oxford: Blackwell.

Smith, M. (2004a). Exploring the implications of the dispositional theory of value. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 297–317). New York: Cambridge University Press. (Reprinted from *Philosophical Issues: Realism and Relativism*, 2002, *12*, 329–347)

Smith, M. (2004b). Internal reasons. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 17–42). New York: Cambridge University Press. (Reprinted from *Philosophy and Phenomenological Research*, 1995, *55*, 109–131)

Smith, M. (2004c). Moral realism. In *Ethics and the a priori: Selected essays*

on moral psychology and meta-ethics (pp. 181–207). New York: Cambridge University Press. (Reprinted from *Blackwell guide to ethical theory*, pp. 15–37, by H. LaFollette, Ed., 2000, Oxford: Blackwell)

Smith, M. (2006). Is that all there is? *The Journal of Ethics*, *10*, 75–106.

Smith, M. (2007). In defence of *Ethics and the A Priori*: A reply to Enoch, Hieronymi, and Tannenbaum. *Philosophical Books*, *48*(2), 136–149.

Sobel, D. (1999). Do the desires of rational agents converge? *Analysis*, *59*(3), 137–147.

Watson, G. (2004). Free action and free will. In *Agency and answerability: Selected essays* (pp. 161–196). New York: Oxford University Press. (Reprinted from *Mind*, 1987, *96*, 145–172)

Williams, B. (1981). Internal and external reasons. In *Moral luck: Philosophical papers 1973-1980* (pp. 101–113). Cambridge: Cambridge University Press. (Reprinted from *Rational action*, by R. Harrison, Ed., 1980, Cambridge: Cambridge University Press)