THEORIA, 2008, **74**, 60–78
doi:10.1111/j.1755-2567.2008.00006.x

Action Explanations Are Not Inherently Normative

by
TILL GRÜNE-YANOFF
University of Helsinki

"Though this be madness, yet there is method in't."
Hamlet, act II, scene ii

Abstract: Inherent normativity is the claim that intentional action explanations necessarily have to comply with normatively understood rationality constraints on the ascribed propositional attitudes. This paper argues against inherent normativity in three steps. First, it presents three examples of actions successfully explained with propositional attitudes, where the ascribed attitudes violate relevant rationality constraints. Second, it argues that the inference rules that systematise propositional attitudes are qualitatively different from rationality constraints both in their justification and their recipients. Third, it rejects additional conditions on propositional attitudes, which purport to necessitate a normative commitment. Thus, inherent normativity is rejected; and with it the claim that intentional action explanations differ substantially from other explanations because they are inherently normative.

Keywords: action explanation, norms of reasoning, rationality, social sciences.

1. Introduction

THEORIES OF INTENTIONAL action explain people's behaviour in terms of their reasons. Some philosophers have claimed that the concept of reason, when employed in explanations, is *necessarily* related to *norms* of reasoning. They argue that intentional action explanations commit the explainer to the normative appropriateness of the ascribed reasoning. In this sense, they claim that intentional action explanations are *inherently normative*. On the basis of this diagnosis, some have concluded that the social sciences must employ fundamentally different forms of explanation from those used in the natural sciences.

To the contrary, I argue that in explaining a person's behaviour, we are not committed to present her as reasoning in a normatively appropriate way, whatever these norms are. All we are committed to is ensuring that the theoretical explanation we give adheres to minimal standards of good scientific practice.

Section 2 presents the claim of inherent normativity and the principle RR that it requires for intentional action explanations. Section 3 discusses three examples of successful intentional action explanations that apparently violate RR. Section 4 proposes an alternative principle IR for intentional action explanations which does not require the portrayal of agents as adhering to norms of rationality. Section 5

argues that IR is different in at least three aspects from RR: in its strength, its addressee and its justification. Section 6 rejects further arguments for RR being necessary for intentional action explanation. Section 7 concludes that inherent normativity is not necessary for intentional action explanation, and that, for that reason, the social sciences are not required to employ forms of explanation fundamentally different from those used in the natural sciences.

2. Inherent Normativity

An agent's intentional action is explained by citing her reasons, (i) if these reasons brought about the action, and (ii) if they make it intelligible how the agent's cognitive and conative constitution led her to act in this way. Most philosophers take an agent's reasons for an action to be a combination of the agent's propositional attitudes – desires, beliefs, values, etc. – which motivates the agent to perform this action. The social sciences often employ such motivating reasons in their explanations.

Frequently, these explanations appeal to principles with a seemingly normative flavour. Take, for example, standard microeconomics models, which strive to explain economic agents' behaviour with reference to their preferences and beliefs: these preferences and beliefs are constrained by *rationality requirements*;[1] and they are connected to the agent's choices via a *maximisation rule*. Many economists think that these requirements and procedures impose normative *constraints* on economic explanation. They admit that the models only explain agents' rational behaviour – irrational behaviour remains unexplained by them. This admission is trivial when taken by itself. But, when seen together with the remarkable resilience of the rational deliberation model in economic research, and combined with economists' long-term neglect of developing models that explain irrational behaviour, it signals economists' conviction that these rationality requirements are crucial to behaviour explanation. Implicitly, one may suspect, they believe that these requirements are not just contingent assumptions of some models of deliberation, but necessary constraints on all models that strive to explain agents' behaviour with reference to their reasons. This is what I call the claim of *inherent normativity*: the conviction that behaviour explanation necessarily has to adhere to normatively understood rationality constraints on the ascribed reasoning and deliberation processes.

1 Typically, these are transitivity and completeness on preferences, and Bayesian probability axioms on beliefs.

Few economists have openly defended inherent normativity.[2] However, eminent philosophers concerned with the methodology of the social sciences have done so. They have claimed, in one form or another, that to explain an agent's behaviour necessitates presenting her reasoning and deliberation as adhering to norms of rationality:

> . . . if we are . . . usefully to describe motions as behaviour, then we are committed to finding, in the pattern of behaviour, belief, and desire, a large degree of rationality and consistency (Davidson, 1980, p. 237) . . . [there is] an irreducibly normative element in all attributions of attitude (p. 241).
>
> . . . propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be (McDowell, 1985, p. 389).
>
> . . . when we are not [rational], the cases defy description in ordinary terms of belief and desire (Dennett, 1987, p. 87) . . . I want to use "rational" as a general-purpose term of cognitive approval (p. 97).

Normatively understood rationality commonly includes judgements about good reasoning as well as good deliberating. It concerns what constitutes valid inference or good argument, as well as what constitutes good and cogent reasons for acting.[3] In this paper, I will argue that this normative rationality requirement differs from what I call the intelligibility requirement, and that only the latter is necessary for action explanations. To appreciate this distinction, we first have to clarify what the above-quoted authors meant by "rational". At least three accounts of rationality can be distinguished in their writings.

First, it has been claimed that rationality is in the eye of the beholder. The interpreter judges whether the interpreted adheres to her reasons and reasoning. Reasons, to be ascribable at all, must necessarily be constrained by "our own standards" (Davidson, 1980, p. 137), "our own reasoning powers" (Davidson, 1987, p. 47), or by the principles of decision theory that are everybody's "own principles" (Davidson, 1985, p. 351). In other words, the essential element of this rationality is "the way I understand my own case when I act for a reason. That is the model for how I understand others" (Schueler, 2003, p. 160). To be of scientific interest, reasoning standards and principles must be shared among interpreters. If standards were highly idiosyncratic, my interpretation of your action would be useless for another person seeking to understand your behaviour.

---

2 It therefore remains open to speculation as to whether their insistence on models of rational behaviour has indeed been motivated by their implicit endorsement of inherent normativity. This paper does not investigate this historical question.

3 It usually does not concern moral norms, or, more generally, prescriptions on which desires or values we should have.

From this consideration, it is but a short step to a stronger position, which defends the idea that rationality requires the adherence to some general principles. Reason ascriptions in intentional action explanations must let the agent appear to be "consistent, a believer of truths, and a lover of the good . . . by our own lights" (Davidson, 1980, p. 222).

Here the agent's behaviour can be explained only if she is seen as having mostly true beliefs, sharing many desires with her fellow humans (like "the desire to find warmth, love, security, and success, and the desire to avoid pain", Davidson, 1982, p. 302), and if her reasoning approximates the rules of (classical) logic, statistic and decision theory.

This requirement would be too strong if it insisted on the unconditional adherence to general principles. It may well be that the belief that should be held rationally by an agent standing in front of a ketch, under the circumstances, is not "That is a ketch", but "That is a yawl". Perhaps the ketch is disguised to look like a yawl. Perhaps the agent's eyesight is failing. Perhaps the agent has been told, by a normally reliable informant, something false – namely, that boats whose jigger is in the position of that of the boat in front of him are yawls, not ketches (Davidson, 1974, p. 196). For this reason, Davidson maintains that we must take agents' circumstances into account when making these ascriptions. It is by the agent's own lights that we judge her – if she has good reasons to form a false belief, or make an invalid inference, then this is not to count against her rationality. Rationality is thus dependent on the process of how the reasons and reasoning procedures were acquired.

Despite these caveats, both accounts fall into the category of *deontologically* justified rationality norms. It is the appeal to principles of their formation, consistency and inference, appropriately conditionalised on the interpreter's perspective or the agent's context, that makes us judge sets of propositional attitudes to be rational or not. This must be distinguished from consequentially justified rationality norms, which appeal to the goals of good reasoning and the efficiency of good deliberation. Typical examples of these accounts include norms based on pragmatic or fitness criteria. Dennett, for example, says that pragmatic rationality constraints must be invoked in intentional action explanations; he insists that explanation must portray the agent in question to adhere to "the proposed (or even universally acclaimed) methods of getting ahead, cognitively, in the world" (Dennett, 1987, p. 97).

In contrast to the deontological accounts, Dennett's proposal allows for two caveats. Firstly, rationality of reasoning operations is judged only against *normal* situations for which the operations were designed. Secondly, these operations must work "most of the time in the contexts in which they are invoked" (Dennett, 1987, p. 96) – so that occasional glitches, malfunctions or mistakes are allowed.

These proposed minimal standards differ substantially. Moreover, they specify neither the principles that are required for rationality, nor what level of success

would be requisite for it. This is prudent, as there is no unique set of principles universally applicable to all reasoning and deliberation. Instead, intuitions about what is rational and what is not often depend on the environment in which the reasoning or deliberation is performed. Furthermore, we often explain people's behaviour with reference to bad arguments, reasons or imprudent plans. To claim that behaviour can only be explained under full ideal rationality is implausible. There must be a certain leeway for violating rationality, within which reason attribution and explanation is still possible.

What unites the three accounts, however, is the imposition of rationality as a *norm* on explanation, and the *justification* of these constraints, by something distinct from the values of good scientific practice. I therefore suggest interpreting all three accounts as advocating the following Rationality Requirement:

RR When explaining actions with propositional attitudes, the ascriptions must make the agent as rational as possible. "Rational" refers to norms of reasoning either deontologically or consequentially justified.

The formulation of RR is deliberatively weak. Those who defend inherent normativity do not spell out just how rational agents have to be. They allow themselves an easy defence of their position by thus being able to claim that any criticism assumes too strong a concept of rationality. The present formulation of RR acknowledges this ambiguity and does not try to specify rationality substantially. Instead, it characterises the rationality required for action explanations as a set of principles that are justified by norms of reasoning, deliberation and action.

This formulation nevertheless has critical content, as it distinguishes RR from other ways in which the concept of rationality is employed in the social sciences. To name but one example, it distinguishes RR from the use of rationality as a heuristic principle. As a heuristic, rationality is a set of convenient commencement assumptions of one's explanatory strategy, which can be gradually weakened or abolished altogether. Consider, for example, the following methodological advice:

> start with a game or naturally occurring situation in which standard game theory makes a bold prediction based on one or two crucial principles; if behavior differs from the prediction, think of plausible explanations for what is observed, and extend formal game theory to incorporate these explanations (Camerer, 1997, pp. 167–168).

In contrast to this, defenders of inherent normativity claim that explanation is impossible, if the minimal norms of rationality are not met.

If rationality were a norm of explanation, justified by values different from values of scientific practice, its purported necessity for intentional action explanations would render these explanations substantially different from those of the natural sciences. According to this, inherent normativity would constitute an argument against a unified concept of scientific explanation. And indeed, all of the

above defenders of inherent normativity have used their claims, in one form or another, in order to drive a wedge between the natural and the social sciences. In the following, I will argue against inherent normativity, and hence against the need for this distinction.

3. Three Examples Against RR

In this section I present three apparently successful intentional action explanations that violate RR.

The first case concerns the explanation of behaviour elicited in Wason and Shapiro's (1971) famous card selection task. They presented subjects in an experiment with four cards laid out on a table, each bearing a single character on the side facing up:

$$A\ B\ 2\ 3$$

Subjects were told that each card has a number on one side and a letter on the other side and that the following conditional statement is true: "If a card has an *A* on one side, then it has an even number on the other side." Subjects then were asked to pick those cards that needed to be turned over in order to figure out whether the conditional statement was indeed true. The most common response was to pick only the card with the *A* on it; only a few checked whether the 3 had a *B* on the back.

This choice cannot be explained by ascribing it to the agents' deductive reasoning capacities in accordance with classical propositional calculus, because they neglect the relevance of *modus tollens*. Instead, a plausible and parsimonious explanation of the majority's behaviour is that they follow *modus ponens*:

$$A, A \rightarrow C \mid\!- C,$$

but neglect *modus tollens*:

$$C, A \rightarrow C \mid\!- \quad A,$$

where , $\rightarrow$ are the standard operators of propositional logic, and $\mid\!-$ is standard Propcal deductive inference. Ascriptions of reasoning characterised in this way violate RR. First, it may incur serious pragmatic disadvantages, as the reasoner is unable to infer potentially important information. Second, it does not adhere to "our standards" – presumably, most of "us" are able to detect the insufficiency of the exhibited reasoning upon brief reflection. Third, it does not adhere to classical propositional logic – the relevant normative principle in the area of qualitative theoretical inferences. Advocates of inherent normativity may rely on this adher-

ence being evaluated by the agent's own lights, as Davidson insisted in the above quote. Accordingly, there may be a story that gives the agent good reasons to neglect *modus tollens*, so that her behaviour is not irrational. If such a story were given, then RR may not be violated.

Crucially, however, such a story is *not* given, and the proposed explanation, based on a weakened classical calculus (which explicitly excludes *modus tollens*), still seems acceptable. Thus, *without* offering an apologetic story for ascribing certain seemingly irrational inference rules to the agent, this ascription offered an explanation. The ascription thus violated RR, and still produced an explanation.

For the second case, imagine an agent who, as all the evidence suggests, is a competent English speaker. In many circumstances, after observing an occurrence of p, she utters a sentence whose meaning in English is that p. Such behaviour may seem to many not only norm-violating, but also unintelligible. However, the latter judgement may change, when we get the following additional information: the agent at those times also expresses her belief that $p \rightarrow q$, and a strong aversion to q. On the basis of this extra information, we can make the agent's behaviour intelligible through the following inference rule.

$$\frac{\text{The agent believes } p \rightarrow q \qquad \text{The agent desires } q}{\text{The agent believes } p}$$

We may call someone whose behaviour can be interpreted as adhering to such an inference a "wishful thinker" (Levin, 1988, pp. 202–203). Again, ascribing such reasoning violates RR. Agents who reason in this way can be easily exploited, as they will in some cases believe states to be false even though they are quite obviously true (and vice versa). The reasoning does not adhere to "our standards" or "our norms" – no standard folk psychology allows inferring a belief from a desire in this way; nor is there any theory of reasoning or decision-making that advocates such a reasoning principle. Again, no apologetic accounts are given, and yet the ascription of the "wishful thinking" inference provides an explanation.[4]

The third case concerns choices based on direct imitation. A particularly striking example is the explanation of local difference in child mortality in villages in southern Germany in the nineteenth and early twentieth century by locally varying practices in breastfeeding. Statistical surveys show that within a village, breast-feeding patterns were fairly homogenous, but between regions, breastfeeding habits varied widely. Between neighbouring regions, the percentage of mothers who never

4 Distinguishing beliefs accepted out of regard for truth from beliefs for the sake of felicity, and excluding the second from the set of belief proper (as Millar, 2004, p. 51 does), in order to rescue the claim that beliefs always are epistemically justified in the eyes of the agent, begs the question.

breastfed – and proportionally, the number of infant deaths – varied by up to 30 per cent (Knodel and van de Walle, 1967, table 5). These differences in customs apparently were long-established traditions, dating back to before the sixteenth century (Knodel and van de Walle, 1967, p. 119). Even though private information was presumably available – in the form of differences in child mortality between villages, observable to the villagers – villagers did not decide on the basis of that information. Rather, they imitated local customs, to the detriment of the children and the sorrow of the parents.

In some cases, following the custom may have been the result of rational deliberation. Knodel and van de Walle report of a woman who moved to Bavaria from northern Germany. Her customary breastfeeding habits were met with social condemnation and threats (p. 129). Under these circumstances, giving in to social pressure may be a rational decision. But the resilience of the habit in the face of available evidence, combined with the rarity of such cases of social pressure, show that the majority of women (in particular if they were raised in the area) never rationally deliberated on whether to follow the prevalent custom. Instead, they directly mimicked their peers. The reasoning implicitly ascribed to them by the authors is therefore something like the following:

$$\frac{\text{The agent believes that relevant agents do } A}{\text{The agent intends } A}$$

Again, the ascription of imitation as a decision procedure violates RR. The increased number of infant deaths clearly speaks of the pragmatic disadvantages of such a practice. Also, "our standards" seem to warn against copycat behaviour and commend instead autonomous, well-informed decisions. Lastly, there are no principles that prescribe direct, blind mimicking. Again, there may be circumlocutory ways to justify imitation, like giving in to peer pressure, or lack of information. The available evidence presented in Knodel and van de Walle, however, does not support any such apologetic stories. Nevertheless, the paper provides an explanation of the varying breastfeeding patterns with reference to imitation, thus violating RR.

4. The Intelligibility Requirement

Action explanations explain by showing that the explanandum – the agent's intentions to act, or propositional attitudes that contribute crucially to the formation of these intentions – fits a certain pattern. The explaining pattern of *intentional* explanations consists of cogent propositional attitudes. Such attitudes – like beliefs, desires, hopes, etc. – are partly specified by their semantic content. Because of their

semantic content, propositional attitudes are subject to consistency constraints. For example, it is meaningless to attribute to an agent both a belief *that A* and a belief *that not A* at the same time.

The propositional attitudes are ascribed to an agent based on observations of her behaviour and environment. As these observations tend to underdetermine the ascriptions, the relations among propositional attitudes, environmental conditions and behaviour are regulated by a theory. This theory creates a "web of propositional attitudes" connected to observations only at its endpoints. If all relevant observations can be subsumed under a set of propositional attitudes without violating any of the theory's constraints, it is correct to say that the observations can be interpreted consistently by that theory.

Theories regulate the relations among propositional attitudes, descriptions of environmental conditions and behaviour by imposing three kinds of inference relations over the attitudes' propositional content: (i) inferences from perceptions to propositional attitudes; (ii) inferences from propositional attitudes to other propositional attitudes; and (iii) inferences from propositional attitudes to intentions. Because these inferences "anchor" the propositional attitudes to each other, to perceptions and to intentions, I will say that they *inferentially characterise* propositional attitudes.[5] Consistent interpretation of an agent's behaviour results in a pattern of propositional attitudes ascribed to the agent. If the semantic content of attitudes in this pattern is *inferentially* related to the explanandum's description, then the pattern explains: it describes how the cognitive and conative constitution of the agent (represented in the propositional attitudes) led to the intention to act.

On the basis of this account, I propose the following Intelligibility Requirement for successful intentional action explanation:

(IR) When explaining actions with propositional attitudes, the ascriptions must make *intelligible* the agent's derivation of the intention to act from her cognitive and conative constitution as represented by the propositional attitudes ascribed to her. "Intelligible" refers to the formulation of a well-behaved inference rule that licenses the derivation.

What is a well-behaved inference rule? In the most general sense, inferences derive conclusions from what is already known. An inference of interest for the present purpose must (i) derive conclusions exclusively from propositional attitudes

5 Actual theories used for interpretative purposes vary widely in the ways that they express such inferential characterisation: qualitative decision theory imposes some kind of logical closure operator; quantitative decision theory in addition posits probability axioms; standard "folk" theories use reasoning schemes. All of them, in one way or another, *make use* of inference rules – logical, statistical, practical and otherwise – to characterise a pattern of propositional attitudes that may subsume the explanandum.

ascribed to the agent; (ii) conclude the explanandum of interest; (iii) be of general form; and (iv) be non-explosive. The first two conditions, I think, do not require further discussion. The third condition is based on the intuition that our reasoning is based on general principles, not particular derivations. Maybe these principles operate only in a highly specific domain. The ascription of reasoning principles that operate only for one particular instance, however, would not have any explanatory value.

The fourth condition is based on the intuition that to be explanatory, inferential characterisation must constrain the ascription of propositional attitudes. Unconstrained ascription does not amount to explanation: if the patterns specified by the inferences subsume *every* relevant or possible configuration of explananda, the explanation is trivial. The exemplary case of such trivial subsumption is the inconsistent set, which – under standard propositional calculus – becomes "explosive": every proposition can be deduced. For example, imagine we wanted to explain an agent's behaviour by reconstructing her theoretical reasoning as adhering to the "tonk" rule. "Tonk" is an absurd operator defined as sharing the introduction rule with "or" and the elimination rule with "and" (Prior, 1960). That is, from p, one infers p-tonk-q and from p-tonk-q one infers q. A theory that ascribed propositional attitudes to an agent on the basis of tonk would ascribe all possible attitudes. Reconstructing an agent's reasoning in this way would hardly amount to an explanation of her behaviour. After all, *any* behaviour could be subsumed under such a propositional attitude "pattern" – which would not so much constitute a pattern but unstructured, uninformative white noise. Thus, ascribed inference rules must be carefully constrained so as not to yield explosive sets of propositional attitudes.

The three examples discussed in section 3 satisfy IR, but do not satisfy RR. In the following section, I will argue that IR and RR are truly distinct, and why adhering to the former does not support the inherent normativity claim.

5. IR Does Not Imply RR

Defenders of inherent normativity may argue that requiring explanations to satisfy IR already implies that these explanations satisfy rationality norms required in RR. This defence fails, because norms of inferential characterisation differ from rational prescriptions on reasoning in at least three ways.

First, IR is weaker than RR. We intuitively distinguish intelligible behaviour from rational or near-rational behaviour. This intuition can be clarified by Michael Smith's distinction between motivating and normative reasons (Smith, 1994). Motivating reasons are psychological states represented by desires and means-ends beliefs governed by inferential constraints. Because they subsume agents' actions under an intelligible constraining pattern, they have the potential to explain.

Normative or justifying reasons, on the other hand, are considerations, or facts, that rationally justify certain actions on an agent's behalf. They are propositions of the form "acting in such-and-such a way in such-and-such circumstances is desirable". Normative reasons commonly also refer to propositional attitudes; hence they also obey inferential characterisation. But to the extent that the conceptual pattern constituted by normative reasons does not subsume agents' actions, they do not explain. Thus motivating and normative reasons are clearly distinct.

Under specific circumstances, however, motivating reasons may *also* be justifying reasons: a desire for an outcome *O* and the belief that action *A* brings about *O* may justify the conclusion that *A* is in fact desirable. This is an *extra* feature of motivating reasons, and not necessary for their role in explanations at all. In the sense that motivating reasons may additionally but contingently assume the justifying role, IR is weaker than RR. Just because we can make sense of someone's behaviour in terms of her motivating reasons, it does not follow that we must conceive of this behaviour as adhering to norms of rationality.

Some defenders of inherent normativity may try to rescue the purported dependency of motivating reasons on justifying reasons by weakening the relevant norms. For example, they may argue that motivating reasons must be the *most* justifying reasons. RR, after all, requires the portrayal of agents as being as rational as *possible*.

Such a reply risks losing the gist of what is normative about norms of rationality. In particular, there are no norms that distinguish between "bad" and "worse" kinds of reasoning. Take, for example, the explanatory claim that an agent systematically makes inferences via affirming the consequent. Is there any norm of rationality which says that such reasoning is better than not reasoning systematically at all? I do not know of one. Rather, it seems to me that this is not a normative question. It is *not* more rational of the agent to reason via affirming the consequent than to reason without any inference rule – both kinds of reasoning violate the RR. The crucial difference, rather, is with respect to explanation: by depicting the agent as reasoning systematically (but norm-violating), the agent's reasoning becomes intelligible and hence explainable, while depicting her reasoning as non-systematic does not explain it at all. Hence, the IR must not be conceived as (approximations of) norms on an agent's action and deliberation, but as general conditions for the possibility of explanation.

Second, while IR and RR both have normative content, the norms underlying RR have a different addressee from those of IR. IR is a prescription for how to perform an intentional action explanation. It prescribes how an explainer ought to construct and use the explanatory theory: "if you want a working theory, you should construct it so that its set of attributed propositional attitudes does not become explosive!" If nobody attempted action explanations, these norms would be without a recipient. This is not the case for the norms of rationality underlying RR: they are directed

towards reasoning and deliberating. They tell reasoning and deliberating agents how they ought to regulate their thought or conduct. As long as there are reasoning agents, these norms have a recipient, independent of whether someone tried to explain their conduct.

Conversely, an agent's conduct may be explained with propositional attitudes even though the agent does not reason in any standard sense (for example, in the case of non-human animals, machines, or even plants) – in which case rationality norms would not have a recipient, while inferential requirements would. Furthermore, as argued in section 3, agents may be subject to norms of rationality and fail them, but their behaviour can still be explained. Thus, theories are subject to norms of inferential requirements, while agents are subject to norms of reasoning and deliberation. To claim that the first implies the second confuses their domain of application.

Third, IR and RR have different justifications. As with other theories, theories of intentional action are used with the goal of describing and systematising present and new data, and investigating which new data may be subsumed under it and how it coheres with other theories from related domains. To be useful for these tasks, theory users must be able "to grasp how the predictions are generated, and to develop a feeling for the consequences the theory has in concrete situations" (De Regt and Dieks, 2005, p. 143). The theory has to be *intelligible*. Intelligibility of theories facilitates the theory's use by allowing a skilful user to apply it to specific phenomena. Because it contributes to the epistemic aims of science in this way, intelligibility is a necessary condition for successful theoretical explanations. In order to be intelligible, as argued in section 4, theories of intentional action must satisfy constraints on inferential characterisation. Otherwise, deductive "explosion" threatens. IR is thus justified by the specific epistemic goals of scientific inquiry.

By contrast, norms of rationality are not derived from epistemic goals, or at least not from epistemic goals *alone*. They are justified either *consequentially* by the goals of good reasoning and the efficiency of good deliberation, or *deontologically* with reference to universal principles of logic, arithmetic and probability calculus. IR and RR differ in their sources of justification, and the claim that inferential constraints constitute rationality constraints in any interesting way confuses these different justifications.

An important implication of the third argument is that IR is mandated by norms of good theory construction which apply to *all* theories. Explanatory theories of the natural sciences must refer to some form of graspable regularity to be intelligible and hence epistemically useful. In theories involving propositional attitudes, inferential characterisations ensure this regularity. Inferential constraints, albeit being norms, thus do not distinguish intentional action theories from other scientific theories. Because there is nothing interestingly normative about regularities in

natural science theories, there is nothing interestingly normative about inferences governing propositional attitudes either.

Thus, the constraints imposed through IR differ from rationality requirements in three ways: they are considerably weaker, they have different domains of application, and they are justified by different sources. Thus, there is no hope of making a case for inherent normativity based on IR.

6. No Rationality Required Besides IR

Defenders of inherent normativity may argue that there is another necessary condition besides IR that is required for intentional explanation, and that this condition requires rationality in a way that supports inherent normativity. I will discuss the most important kinds of these proposed conditions here.

The first argument claims that the norms underlying RR are necessary to deal with underdetermination. Theories of intentional action are notoriously underdetermined. Defenders of inherent normativity suggest that rationality requirements crucially reduce the degrees of freedom. The gap between the evidence, described in terms of behaviour, and the theory, couched in intentional terms, is filled by the assumption that the subject is rational in forming and acting on beliefs and desires. Importantly, defenders of inherent normativity see these assumptions as a *conceptual necessity*, rather than methodological advice. They rule out *in principle* that other behavioural evidence in different but related situations may help determine the weakest assumption in the scheme. Without assuming rationality, they claim, the gap cannot be closed: RR thus becomes the precondition for the explanatory force of reason explanation.

Underdetermination makes it necessary for theorists to start with some heuristic assumptions. It is standard practice to use normatively justified rationality assumptions from decision theory, etc., for this purpose. But as soon as theories built in that way fail, some of the assumptions can be altered. Underdetermination makes this procedure somewhat more complex, but not impossible: methods of bootstrapping have been widely used to get around this problem. Hence RR is not *necessary* for explanation.

Furthermore, the use of heuristic assumptions is nothing specific to theories of intentional action; *all* theories subject to the underdetermination problem use heuristic assumptions in this way. They all supplement the criterion of data fitting and intelligibility with further criteria of elegance, simplicity, plausibility, etc. Their use does not make these criteria a normative requirement for the theory; hence neither does it make RR a normative requirement.

A second argument claims that the norms underlying RR are necessary for the ascription of propositional attitudes. For this, Davidson draws an analogy between

reason ascriptions and numerical measurement. In numerical measurement, numbers are used to index magnitudes of physical objects, such as length or temperature. In reason ascription, semantic content is used to "index" states of agents, such as beliefs or desires. Numerical measurement, as it is well known, requires the measured magnitude to be structured in a particular way. For example, the relation *longer than* must satisfy *transitivity* to be numerically measurable. If it were not the case that *X is longer than Y* and *Y is longer than Z* implies *X is longer than Z*, the relation *longer than* could not be mapped on a numerical ranking. That *longer than* is indeed transitive is secured by a theory of rigid objects. This theory is empirically confirmed by all the phenomena to which it applies. A similar theory, Davidson claims, is required for propositional attitudes:

> Just as one cannot intelligibly assign a length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions (Davidson, 1980, p. 221).

This analogy is notoriously murky, and has drawn many critical responses. What Davidson presumably means is that for propositional attitudes to be meaningfully assigned to an agent, a theory about the agent's reasoning has to secure that it matches the semantic nature of propositional attitudes, just as a theory of rigid objects secures that their length property matches the numerical nature of the measuring scale. Such a theory, Davidson implies, must be based on minimal rationality assumptions. Rationality is thus constitutive of interpretation and hence of intentional explanation.

In this vein, Davidson has defended the transitivity of preferences (1980, p. 273). Violating transitivity, Davidson claims, undermines the very meaning of preferring one option to others, because the relation between preference and choice is lost. A critic, however, can point out possible connections between intransitive preferences and choice. Schwartz (1972), for example, defines a choice function on the basis of intransitive preference orders. Such a choice–preference relation allows systematising choices that violate some of the standardly assumed choice properties, and which therefore are usually considered non-interpretable. Even though transitivity is often considered the bedrock of normative requirements on preferences, it is thus possible to explain agents' behaviour on the basis of intransitive preferences.[6] Hence, at least in some cases, viable theories of norm-violating propositional attitudes exist. A normatively understood rationality therefore cannot be constitutive of interpretation.

---

6 Importantly, this is very different from Davidson's concession of occasional irrationalities. A theory like Schwartz's allows for widespread and systematic violations of transitivity.

A third argument claims that RR is supported by evolutionary considerations. It has been claimed that evolutionary pressure has selected against severely norm-violating reasoning and deliberating: "People who are inconsistent [in their preferences] will necessarily be sometimes wrong and hence will be at a disadvantage compared to those who are always right. And evolution is not kind to memes that inhibit their own replication" (Binmore, 1994, p. 27).

The specific norms of rationality selected for greatly depend on the context in which the reasoning and deliberation effectively determines behaviour. Local violations may occur because an agent behaves in a "safe" environment where certain norm-violating reasoning does not yield any disadvantageous consequences. But, to the extent that ordinary life environments for human agents can be considered similar, this argument leads to the conclusion that in explanations, explainers should impose their own standards on others. The norms of rationality become a piece of evolutionary "design", which must be adhered to in explaining behaviour.

Besides significant doubts about the claim that natural selection always leads to optima, this argument fails because it argues for the acknowledgement of a fact, not a norm. If the evolutionary argument is correct, then it establishes that human agents are rational in some way, and any explanation of human behaviour ought to acknowledge that fact. What does not follow is that explainers must acknowledge the normative appropriateness of behaviour. If something is designed in a certain way, then knowing the design and its purpose is of great importance to the explanation. But whether it *ought* to be this way is wholly irrelevant. There is simply no question of normative appropriateness for the theorist.

A fourth argument claims that RR is implied in the concept of reason. For an agent to have a reason for an action requires that the agent acknowledge the normative force of that reason:

> The question is why she [did what she did], that is what *her reasons* were for [doing it]. So we need to understand her response to this question as somehow describing her reasons. This means that we need to understand her response as *somehow*, at least in her own mind, supporting the thought that she *should* do it, that is as giving considerations *she took to be* reasons for doing it (Schueler, 2003, p. 131; italics in original).

To intentionally explain the agent's behaviour, so the argument goes, one needs to determine not only that the agent adhered to an inferentially characterised pattern, but also to show that the agent acknowledged this normative force – that she took the reasons (rightly or wrongly) to be adequate normative reasons. Without this extra assumption, explanations explain neither with reasons nor do they explain intentional actions: "if we drop the attribution of some evaluative premise . . . to the practical reasoning of the agent performing the action, we are thereby dropping . . . the thought that this person acted for a reason at all, and hence the thought

that the agent intentionally acted at all" (Schueler, 2003, p. 135). Because the concept of reason commits us to making the agent's normative view plausible, reason explanations "are always implicitly normative in at least the minimal sense that they necessarily attribute to the agent a normative view (about what she should do) based on evaluative premises" (Schueler, 2003, p. 149).

Schueler's conclusion depends on two questionable presuppositions. First, we do not always answer the question of why someone did what she did by identifying the rules that she *followed*, but only the rules she *adhered* to in her behaviour. Agents often act on reasons of which they are not consciously aware. Hans, for example, may exhibit certain table manners, without being aware that his reason for doing so is his desire to appear sophisticated; he just has the disposition to do so, as a result of his parents educating him in a certain way. To ascribe to Hans the *motivating* reason that he wants to appear sophisticated may be a good explanation of his behaviour, especially if he exhibits other behavioural patterns (say, a certain mannerism in his expressions, or rather extravagant interests) that could be explained by the same reason. Depending on other, maybe more explicitly held, justifying reasons, Hans's desire to appear sophisticated may also be a *potentially* justifying reason for his table manners. But it cannot be an *actual* justifying reason, because he never justifies his table manners in this way. He does not go through an explicit practical reasoning scheme, deriving from his desire to appear sophisticated that he should exhibit certain table manners – he just has them. Nevertheless, I submit, we would still consider Hans's eating in a certain fashion an intentional action. So even though we, as observers, note his motivating reasons (and may snicker at his pretensions), he is unaware of what justifies his behaviour for himself. Hence the connection between giving a motivating reason for why an agent acted, and the agent's taking this as his justifying reason – as the reason for why he should do what he did – are not as close as Schueler wants us to believe.

Schueler's conclusion further presupposes that someone thinking that she should perform an action, because she has a good reason to, imposes a normative constraint on that action's explanation. I disagree. That someone thinks she has a good reason, or evaluates certain reasons as being more important than others, is a fact about that agent. There is no connection between thinking that *R* is a good reason and *R* being a good reason, the latter being a normative claim. Further, that an agent thinks she should (has adequate reason to) perform an action does not imply that she will perform it, unless we have an empirically supported theory of some sort, which establishes that agents tend to do what they think they should do. A reason is explanatory only because there is a correlated fact about the agent – such as that she believes that reasons of this sort are sufficient reasons for action, and that she has a disposition to act in a certain way under those circumstances. Without this fact, the reason is not explanatory, because it does not effectively determinate the action. In the presence of this fact, however, the normative evaluations that the

agent associates with these reasons collapse into a description of the agent's beliefs or dispositions. The normative conclusions of the agent's practical reasoning are therefore irrelevant for the explanatory purpose (Henderson, 2002).

A fifth argument makes a similar claim about the conceptual relation between ascribing reasons and RR, but draws its intuitions from "first person scruples". Looking at ourselves, we cannot avoid the intuition that we as humans have a capacity for deliberative thinking. We think that we have a specific relationship to our intentions and our expectations of how we will carry them out:

> In regarding myself as currently intending to Φ . . . I view my Φing as something that, in and of itself, enjoins me either to do what is necessary or to give up the intention . . . I think of myself as the agent of my Φing . . . I do not view myself simply as one who, because I have certain dispositions and tendencies, is likely to do certain things (Millar, 2004, p. 136).

If we pried explanatory reasons from norms of rationality, we replace the commitment we have to our intentions with a non-normative regularity that only captures a tendency to act. But that, so the argument goes, would eliminate the realm of the intentional altogether. Why talk of "reasons" at all, if all it takes for me to have a reason is to adhere to an inferentially characterised pattern? Why ascribe reasons if there is no deliberative thinking, no agency of intending?

This criticism goes amiss, because it illegitimately generalises the first-person point of view. From that perspective, regarding my current own intending is not (only) an explanation of myself; it (also) involves practical reasoning. That latter practice is clearly normative. But other perspectives – in particular a third-person perspective, but arguably also a first-person perspective on one's past actions – do not necessarily involve such a normative practice. While we judge ourselves when thinking about our reasons for contemporary and future actions, we are able to explain why people decided as they did (or how we decided as we did) without making such a judgement. However, the reasons by which we explain actions are always *capable* of being subject to normative evaluation. So when we explain others' behaviour intentionally, we *in addition* may form a judgement about the rationality of their reasoning and deliberation. The explanation itself, though, does not require us to make that judgement. The realm of the intentional is thus not eliminated by my argument; however, it is not demarcated by the explanatory use of reasons, but by normative practices pertaining to these reasons – and these two uses can be handled separately, thus undermining the need for RR.

A sixth argument relies on the "special interest" that we have in reason explanation. For example, Davidson claims that our special interest in reason explanation is to interpret human agents as persons. He warns that "to the extent that we fail to discover a coherent and plausible pattern of attitudes and actions of others we simply forgo the chance of treating them as persons" (Davidson, 1980, p. 222).

But this special interest – of seeing human agents as rational or as persons – cannot be a reason to exclude other forms of reasoning for explanatory purposes, as long as those explanatory purposes are governed by epistemic goals alone. Maybe there are important ethical reasons not to pursue explanations that let human agents appear as violating basic norms of rationality – as there may have been reasons not to further research into nuclear fission or the human genetic code. But when we decide not to continue this line of research, because it endangers our understanding of others as persons, then we make scientific practice subject to other goals than epistemic ones, and that goes beyond the scope of the present debate.

Thus, although reasons play a role in both explanatory and normative practices, invoking reasons in explanatory practices does not require these explanations to adhere to any norms of rationality. Intentional action theories exploit the intentional idiom, but their fundamental approach is the same as that of other scientific theories, disregarding any special interest concerning the object to be explained.

6. Conclusion

Reason explanations are based on inferentially characterised patterns of propositional attitudes. They are successful if the explananda can be fitted under the explaining pattern. To function properly, the inferential characterisations must satisfy certain minimal requirements; the chief requirement being that they do not deteriorate into explosive inferences.

The IR requirement – that intentional action explanations be intelligible, and therefore be based on well-behaved inferential rules – is derived from the epistemic goals that govern all scientific explanations, and it is directed at the explainer. It is qualitatively different from norms of rationality required by RR, which are stronger, and have different justifications and different addressees.

Propositional attitudes constrained by IR allow explaining behaviour on the basis of RR-violating inferences. Furthermore, RR is not necessary to defuse the underdetermination problem, or to attribute propositional attitudes. It does not flow from evolutionary considerations. Nor is it in any relevant way implied by the notion of reason, or the notion of intentionality. I therefore conclude that in explaining behaviour we do not have to acknowledge RR. Intentional action explanations are not inherently normative.

Of course, it may be methodological good practice to start with the hypothesis that an agent, whose behaviour we want to explain, is rational, or shares most of our desires and beliefs. But if the evidence of her behaviour speaks against these initial assumptions, then a good explanation may require a scheme that proposes reasons characterised by norm-violating inferences. Just like Polonius, who found Hamlet mad, but not without method, so we can explain people's behaviour by showing that

it adheres to some inferentially characterised set of propositional attitudes. The claim that social sciences must employ a radically different form of explanation from those used in the natural sciences therefore is unwarranted, at least with regard to the inherent normativity of intentional action explanations.

Acknowledgements

I thank Mark LeBar, Joanne Grüne-Yanoff, seminar groups at Lund University and Stockholm University, and two anonymous referees for helpful comments.

References

BINMORE, K. (1994) *Game Theory and the Social Contract. Volume I: Just Playing*. Cambridge, MA: MIT Press.

CAMERER, C. (1997) "Progress in Behavioral Game Theory." *Journal of Economic Perspectives*, 11: 167–188.

DAVIDSON, D. (1974) "On the Very Idea of a Conceptual Scheme." In D. Davidson, *Essays on Truth and Interpretation*, pp. 184–198. Oxford: Oxford University Press.

DAVIDSON, D. (1980) *Essays on Actions and Events*. Oxford: Oxford University Press.

DAVIDSON, D. (1982) "Paradoxes of Irrationality." In R. Wolheim and J. Hopkins (eds), *Philosophical Essays on Freud*, pp. 289–305. Cambridge: Cambridge University Press.

DAVIDSON, D. (1985) "Incoherence and Irrationality." *Dialectica*, 39(4): 345–354.

DAVIDSON, D. (1987) "Problems in the Explanation of Action." In P. Pettit, R. Sylvan and J. Norman (eds), *Metaphysics and Morality*, pp. 35–49. Oxford: Blackwell.

DENNETT, D. (1987) "Making Sense of Ourselves." In D. Dennett, *The Intentional Stance*, pp. 83–102. Cambridge, MA: MIT Press.

DE REGT, H. and DIEKS, D. (2005) "A Contextual Approach to Scientific Understanding." *Synthese*, 144(1): 137–170.

HENDERSON, D. (2002) "Norms, Normative Principles and Explanation." *Philosophy of the Social Sciences*, 32(3): 329–364.

KNODEL, J. and VAN DE WALLE, E. (1967) "Breast Feeding, Fertility and Infant Mortality: an Analysis of Some Early German Data." *Population Studies*, 21(2): 109–131.

LEVIN, J. (1988) "Must Reasons be Rational?" *Philosophy of Science*, 55(2): 199–217.

MCDOWELL, J. (1985) "Functionalism and Anomalous Monism." In B. McLaughlin and E. Lepore (eds), *Actions and Events*, pp. 687–694. Oxford: Blackwell.

MILLAR, A. (2004) *Understanding People*. Oxford: Oxford University Press.

PRIOR, A. N. (1960) "The Runabout Inference Ticket." *Analysis*, 21(2): 38–39.

SCHUELER, F. G. (2003) *Reasons and Purposes. Human Rationality and the Teleological Explanation of Action*. Oxford: Oxford University Press.

SCHWARTZ, T. (1972) "Rationality and the Myth of the Maximum." *Nous*, 6: 97–117.

SMITH, M. (1994) *The Moral Problem*. Oxford: Blackwell.

WASON, P. C. and SHAPIRO, D. (1971) "Natural and contrived experience in a reasoning problem." *Quarterly Journal of Experimental Psychology*, 23: 63–71.