Rationally choosing to follow norms

Introduction

It seems very plausible that, at least in some occasions, an explanation of group behaviour will be improved if we are able to give an account of the relevant individuals' behaviour. This gives us reason to look at the various options for laying such "micro-foundations". RCT has emerged as a leading contender as a model of individuals' behaviour. It is particularly attractive to people who want to theorise about society in as scientific a way as possible: RCT often allows a certain amount of quantitativity, and even where it does not, it can still yield predictions that can be empirically tested. Also, some theorists admire the finality of RCT explanations. It seems that once we know that an action was rational for an agent to perform, we need ask no further questions about why she performed it: as Hollis puts it, "rational choice is its own explanation" (Hollis 1977).

But how widely can RCT be applied? One popular criticism is that RCT will often fail to provide adequate explanations of norm-based behaviour. This is significant, because much of our everyday life is structured by social norms: our daily decisions to wear clothes to work, avoid spitting during lunch, to tell the truth, and countless more, depend on social norms. Typically, it is thought that RCT will only be able to explain this behaviour insofar

as it is motivated by an aversion to negative consequences that arise from failing to conform to the norm.

One type of negative consequence is the social sanctioning that can arise from disobeying a norm. Disapproval in itself is unpleasant, and it can also lead to being punished in other ways, and so it seems rational agents will try to avoid being sanctioned. An alternative is to model individuals as facing various "psychic costs", such as those arising from the shame of disobedience.

Elizabeth Anderson takes such a view: "Rational choice theory uses the model of Homo Economicus. It explains behavior in conformity with social norms as the product of the strategic interactions of instrumentally rational, self-interested individuals… Rational choice theory represents individuals as taking a more alienated posture toward social norms. Although they may see that general conformity to a norm would be desirable, this does not provide them with a reason to conform, so long as personal conformity is, on net, costly to each agent. Only incentives contingently attached to the norm could provide a rational, self-interested individual a reason to conform. A person's reasons for conformity are thus *external* to the normativity of the norm, incidental to whatever might make its adherents approve of a general conformity to it." (Anderson 2000).

But these explanations often don't ring true to the phenomenology of norm-based behaviour: often it seems to us that we conform with norms, without regard for any

possible sanctioning or looking forward to how disobedience might make us feel. For example, when I take my place in a queue, and don't cut in front, it seems to me that I haven't calculated about what bad might happen if I don't; rather, it seems that I've thought that this is the appropriate thing to do, and haven't considered the consequences one way or another. This means that these explanations amount to claiming that we regularly engage in self-deception: although it seems to us that we abide by norms for their own sake, this appearance is a mere rationalization, and our real motives are subconscious attractions to the beneficial consequences of conformity, either social or psychic. This would be a bold empirical claim that can't be ruled out a priori, but it is hardly attractive, and we'd do well to look for alternative accounts.

This seems to leave the RCT theorist in a bind: either she must allow her theory does not apply to these, presumably widespread, cases, or she must resort to an error theory about the phenomenology of norm-based behaviour.

In this paper, I will argue that, in fact, there is no such bind, and the appearance of one arises from a confusion between the RCT framework, and particular substantive theories posited within the framework.

Some particular theories model individuals as having selfish preferences for the effects of their actions. These theories face the aforementioned bind. But I will argue that the RCT framework is not committed to any particular theory about an agent's preferences. The

illusion that it is arises only if we fail to be clear about the notions of preferences and utility that lie at the heart of the RCT approach. When we have straightened these out, it will be clear that there is no problem of talking about an agent "rationally choosing to follow a norm for its own sake".

The Target Arguments

It will help to put the motivation for the view that RCT can't explain this norm-based behaiour more sharply. The following picture of RCT is quite common: "RCT models self-interested agents as instrumentally rational: they engage in analyzing outcomes in terms of the costs and benefits that the action is a means to. The model predicts that the agent will choose that action which is a means to the outcome that is most in the agent's interest."

Let us define "norm-driven behaviour" as behaviour performed for its own sake, as a result of internalizing a norm. On the basis of the above picture, the following two arguments suggest themselves, which I will refer to as the "A argument" and "B argument" respectively:

A1. RCT can only explain outcome-oriented behaviour.

A2. Norm-driven behaviour is not outcome-oriented.

A3. Therefore, RCT cannot explain norm-driven behaviour.

B1. RCT models agents as self-interested.

B2. Norm-driven behaviour is not self-interested.

B3. Therefore, RCT cannot explain norm-driven behaviour.

Notice that these arguments are perfectly general and would give us reason to rule out all RCT explanations of norm-driven behaviour. Also, we wouldn't have to look at any evidence or data; we'd just have to know what an RCT explanation is, and know what norm-driven behaviour is, and we'd be able to conclude that RCT can't explain it. In this respect, the arguments present a priori objections to RCT's ability to explain norm-driven behaviour.

I will argue that neither argument is a good one. They are either invalid or have false premises. If the premises in the arguments are true, then there is an equivocation on the terms "outcome-oriented" and "self-interested", respectively, and so the arguments are invalid. If there is no equivocation on these terms, then at least one premise of each argument is false, and so the arguments are unsound. Therefore, both arguments are unsound. By a close perusal of the RCT framework, we will see why this is the case, and thereby understand to what extent the "picture" outlined above is accurate. Although, for completeness, I will sketch all of the essential components of any RCT model, the preferences component will be the most important aspect for our purposes.

The motivations of rational actors

My argument will rely on distinguishing what is essential to the RCT framework from contingent features of theories posited in it. Therefore, we need to know what is essential to the RCT framework. A helpful way into this problem is suggested by Jon Elster (Elster 1985), following Donald Davidson's exposition of a Humean explanation theory of motivation (Davidson 1963). Davidson argued that "rationalisation", or a rational explanation of an action, involved giving the agent's "primary reason" for an action. Primary reasons were belief – "desire" pairs, such that the belief is of the form that an action has a property, and the "desire" representing any action that has this property as attractive. For example, one might explain Dave's eating a banana, by citing his desire to be healthy, and his belief that eating the banana is conducive to being healthy. By giving such an explanation, Davidson claimed we are able to present "what the agent saw in the action".

This places one important constraint on what can count as a desire, though. In order for an explanation in terms of desires to present what the agent saw as favourable in the action, it must be the case that these desires are introspectively accessible to the agent. Subconscious motivations cannot be offered as what an agent takes to be her reasons. In this respect, the Davidsonian approach to rational explanation, nicely fits a distinction

within sociological explanations between those that do and do not offer explanations of actions in terms that the participants would endorse.

This qualification aside, the category of desire here must be understood very broadly to include all sorts of motivation, and some people prefer the term "pro attitude" to mark how this is an explicitly theoretical concept being employed. In grouping them together, the Humean approach is not committed to the claim that all pro attitudes are a homogeneous kind, which, for example, have a shared phenomenological feel. Rather, the motivation for this grouping, as with most explicit conceptual regimentation, lies in what one can do with it. In Davidson's project the grouping enabled him to define and argue for the thesis that rational explanations are causal explanations[1]. In RCT theory, the grouping marks out where in the model of the individual the motivational component of action lies.

This approach is a good start, but as yet, it can't help us with explaining choice: nothing has been said as to why one action was chosen over another. Suppose, following our previous example, as well as knowing that Dave wants to be healthy, and believes eating the banana is good for this, he also likes sweet things, and believes that candy is sweet. If this was all the information we had to go on, we would be unable to predict that he would eat the banana rather than the candy, and so citing this information cannot be a proper

[1] Davidson's main argument was that only a causal approach could make sense of the 'because' in, e.g., "Dave ate the banana *because* he believed it was nutritious". Theorists who want to distinguish intentional from causal explanations (c.f. Cohen ****) nearly universally fail to discharge the burden of explaining what is wrong with this compelling argument.

explanation of this choice. The natural extension to this is to take a comparative view of the various desires an agent has. We need to know which of an agent's various desires are stronger than others. Once we learn that Dave's desire for health is stronger than that for sweet things, we can understand why he chose the banana. In this way, we arrive at one conception of 'preferences'[2]: in our example, we would say that Dave has a preference for being healthy over sweet things. This proposal does require that, if desires are not homogenous, they can still be compared in motivational strength. But, this seems entirely possible. We quite naturally say that one's motivation to do the right thing was stronger than one's temptation to succumb to pleasure, and these types of motivation would be as good contenders for being different types of motivation, as one could expect.

This conception of preferences is perfectly suited to explaining action in terms of an agent's rational choices. It is a motivational conception of preference, and so is what is needed for explaining action. (A conception of preference that was not tied to motivation would not be up to the task at hand). On the other hand, it is a conception in terms of the reasons for an agent's action. This makes it well suited to casting these choices as rational: they are the agents choices when she acts on reasons. This is exactly the conceptual foundation that RCT needs.

---

[2]"x wants to V more than to W" is not the only way of glossing "x prefers V-ing to W-ing" Another would be to understand preferences as normative beliefs. For example, "x believes V-ing is more desirable than W-ing". However, if beliefs and motivations are modally separable, then this conception of 'preference' seems entirely unsuited to the task at hand: citing this belief of the agent would not yet ensure the agent is appropriately motivated to act.

Out of the concept of preference-satisfaction, RCT constructs a notion of utility: an agent's utility consists in the satisfaction of her preferences. Insofar as an action satisfies an agent's preferences, it raises the agent's utility. So, if Fred prefers fishing to tennis, then going fishing yields him more utility. If an agent's preferences are "well-formed", for example by being consistent, we can produce mathematical utility *functions* to represent the agent's preferences. These functions assign numerical values to each option, according to how the agent prefers it. If we merely know which options are preferred to others, we can only produce an ordinal utility function. If we know how much each option is preferred to others, we can produce a cardinal utility function. Now, the numbers the value assigns mean something more: if says how much each is preferred.

It is essential to note what this concept of utility is not. It is not the concept of the agent's happiness or well-being. To say that one option gives an agent more utility than another is not to say that this option will make him better off or happier. Rather, it is just to say that an agent prefers this option. And there could be many reasons why the agent prefers this option. Perhaps, the agent is an altruistic person who can sacrifice themselves by jumping on a grenade to save 100 people in the room. In such a case, the agent may have quite a strong preference for this option. If this is so, it is quite correct to say this option gives him much utility. But, this is not yet to say that the option makes him happy or better off[3].

[3] There are some theories of well-being which link well-being and preferences. According to desire-satisfaction theories, an agent is better off insofar as his desires are satisfied. According to these theories, although it will lead to his death, our grenade-thwarter has made himself better off, at least in one respect. Many people reject these theories of well-being. But even if they are correct, it is important to note that they are making substantive claims. They are not presenting their theory as a triviality that follows from how they set definitions up. For this to be the case, they are committed to the concepts of well-being and preference-satisfaction being distinct concepts.

So, there is nothing in the concept of preference that limits it to egoistic and selfish behaviour[4]. Since the concept of utility that rational choice theory uses is merely a construct from preference, it follows that utility is not limited to egoistic and selfish behaviour. It is often said that RCT involves modeling agents as engaging in "cost-benefit analysis". The only sense in which this is true would be when we are considering benefits to utility. No other notion of well-being, or any normative notion, is involved, out of which one could talk of benefits. Admittedly, this language of benefits is somewhat misleading: it seems strange to talk of the grenade martyr as choosing to do so because this action had most benefits and fewest costs. But once we realise that this merely means he preferred the option, the strangeness disappears. Arguably, the conceptual foundations of RCT would be less open to misconstrual if a more appropriate terminology were used. However, like it or not, the language of 'utility', 'cost-benefit analysis' and the like is

---

[4] There are formal limitations of *some* RCT explanations that may prevent them from modeling certain types of altruistic behaviour. Some models, particularly those using game theory, assume that all agents utility functions can be stated independently of each other. It is a consequence of this that these models cannot handle agents having preferences about each other's preferences. These agents are modeled as acting in accord with their own preferences, and no one elses. In this sense, some RCT models agents as failing to pay attention to certain features of others But, these models can still be used for altruistic behaviour. Consider the following:

(1) Jones gives Bloggs a banana because Bloggs likes bananas.
(2) Jones gives Bloggs a banana because Jones believes Bloggs is a human and that bananas are good for humans

The models under consideration cannot handle behaviour like that in (1), but they can handle (2). In (1) Jones has a preference for giving bananas because Bloggs has a certain preference (for receiving bananas). In (2), Jones has an altruistic preference which is not based on the preferences of any other agent.

Still, one might worry that a large swathe of altruistic action has been left out. Often we want to help others, but the way in which we do is based on what they want for themselves. For example, parents may give children Power Rangers because they know the children want them. If they preferred another toy, they would have given them something else. I take it that the inability to handle these cases of altruistic action are indeed real limitations on these models. However, not all RCT models require that all individuals' preferences can be stated independently, and it's hard to see why this requirement would be made outside of game theory.

entrenched; the best we can do is to be scrupulous about the senses in which we are using these terms.

The important point in all of this is that the RCT framework is entirely neutral as to what an agents' preferences are. For any substantive account of an agent's preferences, the framework confers on these preferences a certain theoretical role, and provides a procedure for deriving from these preferences other theoretical variables (e.g. the agent's utility). But the framework itself is not committed to any particular account of what these preferences are. To be sure, each theory posited in the framework will give a substantive account, but the framework places no constraints on the content of these.

This will be the most important point of this section, but for completeness, a little bit more has to be added to capture all of the framework. As well as involving the cognitive and conative elements of an agent in her beliefs and preferences, the model must take account of the fact that her action is constrained by possibility. Therefore, the model must incorporate the actual situation she faces. For example, suppose Dave falsely believes the hologram is a banana. This fact of the situation will mean he is unable to eat a banana. The last thing the model needs is a "decision-procedure", which would be a function from ordered triples of the form <set of preferences, set of beliefs, situation> to actions. Roughly, this part would model the agent as performing the action that best satisfies the agent's preferences in light of her beliefs, and is actually possible given the situation.

Although it will often be left implicit, this is the structure that is essential to any RCT explanation. This variation can be understood in terms of how one restricts which sets of beliefs, sets of preferences, and actions are admissible, and what the precise detail of the decision-procedure function is. For example, some theorists only admit certain sets of preferences that are "well-behaved", for example by being consistent with each other, on the grounds that only these are "rational" sets to have[5]. These debates will be entirely orthogonal to our interests: how they are resolved has no implications for the content of agents' preferences. They merely place formal limitations n which preferences can be held together, regardless of their content. For example, the requirement of transitivity would require that if an agent prefers A to B, and B to C, then she must prefer A to C. Clearly, such a requirement has no consequences on how A, B and C are interpreted. Another point of dispute is that some models will allow "gather more information" as a possible action to be performed[6], but again this need not concern us, as it has no consequences for preferences.

Now, we have seen what is essential to the RCT framework, we are in a position to return to the framing Target Arguments, and see what is wrong with each. In particular, this will be aided by the previous clarification of the notion of preferences, and the concepts that are defined in terms of this.

---

[5] Some theorists have argued that further constraints must be placed on the preferences and beliefs the agent can have in order for the agent to deserve the name of 'rational'. In the first place, such a dispute is theoretically uninteresting; what we should be concerned with is whether these explanations are good ones, and not what to label them. But, even in the terminological dispute, these theorists have overlooked the Davidsonian point that these choices are seen as rational, insofar as the agents are responding to the reasons they take themselves to have. Whether the agent's conception of her reasons is a coherent one is orthogonal to this point.

[6] When seen in this light, "satisficing theory"

Returning to the target arguments

The moral of the last section is that the RCT framework is committed to modeling agents as preferences, but the framework is not committed to any particular account of what these preferences are. In other words, the framework requires that any theory give an account of these preferences, but places no limit on what this account may be. As such, there is no a priori theoretical reason why these preferences can't be for the performance of actions themselves. Also, there is no good reason why these preferences have to be egoistically focused on the agent's well-being.

With this in mind, we can state precisely in what sense RCT is "outcome-oriented" and "self-interested".

What an agent sees as an "outcome" of an action can be understood in two senses: a *teleological* sense, and a *causal* sense. In the teleological sense, an action's outcome is the goal the agent had in acting. It is what the agent aims to bring about with the action - the point of acting in the first place. In the causal sense, an action's outcomes are the effects it produces. As effects, they must be logically distinct from the action.

However, these two senses come apart. Suppose I return a loan in order to do my duty. Doing my duty would be my goal, and so my outcome in the teleological sense. But, we shouldn't think of the returning of the loan as *causing* the doing of the duty. They are not separate events, such that one caused the other. Rather, the returning *is* the doing[7]. Suppose that, having already decided to return the money, I foresee that doing so will cause me to be considered more trustworthy. This would be an outcome in the causal sense, but it is not my goal when acting, and so is not an outcome in the teleological sense.

So, in what sense is RCT "outcome-oriented"? What RCT explanations must give is an account of the agent's preferences, in order to state the reason why the agent acted. This makes it clear that what it is tied to is the teleological sense. But there is no reason it must be tied to the causal sense. An agent can have preferences for intrinsic properties of the actions themselves, above and beyond its causal powers. In the previous example, we can perfectly well talk of my preference for doing my duty, even though this is not an effect.

However, when people point out that norm-driven behaviour is not "outcome-oriented", they are using the causal sense of "outcome". People often conform with norms for their own sake, and not for any effects distinct from this conformity. So, we can see that the A argument invites an equivocation between these two senses. If we stick to one of these senses, then either A1 or A2 is false. Only by equivocating between these senses, can both premises be true, but in this case the argument is invalid, and A3 does not follow from the

[7] An alternative view of action would hold that the doing one's duty and the returning are separate actions, such that the former is performed by performing the latter. Still, even on this view, the "...by..." relation is not plausibly a causal relation. More likely, it is envisaged as a "part-whole" relation.

conjunction of A1 & A2. Therefore, either the A argument has false premises, or it is invalid. Therefore, the A argument is unsound.

Similarly, we can distinguish two senses in which actions can be "self-interested": the *egoist* sense and the *preference-based* sense. In the egoist sense, an agent acts self-interestedly insofar as she acts selfishly. She does this when she pursues what she sees as her own good, and without regard for the well-being of others. In the preference-based sense, an agent's self-interest is whatever constitutes the satisfaction of her preferences. Therefore, in this sense, an agent acts self-interestedly when she acts in light of her own preferences. Suppose Mary prefers to give up her seat for a stranger. In the preference-based sense, this altruistic action is in Mary's self-interest. This is not a particularly interesting sense: it is tautologous that agents who consciously act because of their desires act self-interestedly in this sense.

It should be clear by now that RCT is only committed to modeling agents as "self-interested" in the preference-based sense. Again, all RCT is committed to is giving an account of an agent's preferences, and there is nothing in the RCT framework that commits one to giving an egoistical account of agents' preferences.

However, when people point out that norm-driven behaviour is not self-interested, they must be using the egoist sense, if what they say is to be true. Therefore, if we stick to one of these senses, then either B1 or B2 is false. Only by equivocating between these senses, can

both premises be true, but in this case the argument is unsound and B3 does not follow from the conjunction of B1 & B2. Therefore, the B argument is unsound.

Why have theorists been tempted to make these equivocations? One main temptation to think this comes from actual RCT practice. As a matter of fact, RCT has been used predominantly to explain egoistic behaviour focused on (causal) effects, and has been most successful when it has. As a result, it is easy to blur the lines between what is widespread in, and what is essential to, RCT.

RCT's birthplace was the field of economics, and in microeconomics it is used to model the behaviour of consumers and firms. In market behaviour, the preferences agents have are to maximize their own profits, income, consumption or some similar property of theirs. RCT models them as motivated about certain results that will come about by undertaking certain courses of action. These agents have only an instrumental interest in the action itself. What they care about is what it is a means to. For example, firms have no intrinsic interest in the price they set. All they care about is that it is the price which leads to them ending up with the most profits. Also, in these contexts, agents are modeled as acting egoistically: they are concerned with their own financial outcomes and not others.

So, economists usually model agents as egoists and (causally) outcome-oriented. In fact, these features are also present in many other areas where RCT has been successful. But we should notice that this aspect of the economic model is a contingent one. It is not forced

on us by the RCT framework itself. Instead, the decision to attribute these selfish preferences to economic agents is empirically motivated. It must be plausible given what we know about human psychology, and it must be confirmed by actual behaviour. If economic agents did not just act to maximize their profits, but helped others as well, this would not show that RCT is the wrong approach in economics. Rather, it would only show that one particular type of RCT explanation is wrong – namely those that posit a particular type of egoistical utility function.

This is not a remote possibility unconnected to actual practice in economics. Work in the growing field of behavioral economics aims to come up with different utility functions that are more realistic, and better account for experimental data. For example, experimental studies find that people prefer to reject deals rather than accept those they perceive as unfair, even though this leaves them at a loss in terms of their assets. A promising response is to make agents' preferences, and hence utility, not merely a function of what happens to them, but also what happens to others.

Applications for norms

We have seen why the arguments for why RCT cannot explain norm-based behaviour are bad ones. It is not a short step to see how it can actually explain this behaviour: we simply impute preferences to agents for conformity with the norm for it's own sake. Since the

ability of RCT to explain sanction-motivated norm conformity is not disputed, let's just consider norm-driven behaviour - behaviour performed for its own sake, as a result of internalizing a norm. Internalising a norm creates certain dispositions within an agent. The relevant one for our purposes is that an agent now desires to conform with the norm. (using the previously mentioned broad sense of "desire"). As a result, the agent will prefer to conform over other actions. The extent of this preference will depend on how deeply internalized the norm is, and the strength of the agent's other desires. Again, it is important to remember that the operative concepts of utility, costs, benefits, and self-interest are all to be defined in terms of preferences. If one may have preferences for conforming with a norm, then it has to be the case that it is proper to talk of the utility, benefit and (preference-based) self-interestedness of doing so. Suppose Laura prefers to leave tips at restaurants, merely to conform with the social norm of doing so. Given this preference, tipping constitutes part of Laura's utility. This means that tipping raises Laura's utility and so benefits her and is in her self-interest.

So, if norm-driven behaviour is amenable to RCT explanation, does this mean that the traditional debate between the asset-maximising *homo economicus* and the rule-following *homo sociologicus* was misconstrued? In one sense, it was, insofar as participants took it to be a debate about whether RCT is appropriate. However, even when we have realized this isn't the case, it doesn't follow that there was nothing to the dispute. There is substantive and interesting disagreement, but this is better understood over which utility functions and preferences we should impute to agents: should we model them as having preferences

based on their (causal) outcomes in terms of their (egoistical) self-interest, or preferences for the intrinsic performance of some actions? In other words, I am proposing that this debate be understood as an internal dispute between which theories we should posit within the RCT framework, and is not about whether the RCT framework can be used to explain certain types of action.

One worry with the current proposal will no doubt be that too much explanatory work has been pushed into explaining the preferences themselves. Take gender-based clothing for example. To 'explain' this behaviour by claiming the agents have preferences to wear these clothes seems to just miss the point. What we are really interested in are where these preferences have come from.

This objection is completely right, insofar as it goes. Any particular RCT explanation cannot explain the components its model uses (i.e., the preferences, beliefs etc), and if we are unsatisfied with the explanation unless some light is shed on these, the RCT explanation is not enough by itself.

However, it need not always be the case that we need the preferences explained. This certainly seems to be so in the case of egoistical behaviour: we do not send back economic models as incomplete on the grounds that they do not include an explanation of why it is individuals prefer more money to less. Moreover, there seems no good a priori reason to always demand this even in the case of norm-driven behaviour. Often, we feel sufficiently

comfortable with the existence of a norm, that we do not cry out for its explanation. We accept that it is empirically well confirmed that people are subject to the norm, and for some explanatory projects, this is good enough. For example, suppose we have good evidence that people are moved by considerations of fairness, and we have good evidence for their beliefs about what is fair. In many cases, we won't be interested in the explanation of why people are moved by fairness, or why they have this conception of what's fair. Rather, we can build these features of people into our explanations, in order to model their interactions in bargaining, without having to give an account of their preferences for conforming with norms of fairness.

The issue of our knowledge of preferences, and any need for them to be explained, presents a natural limit for how and when RCT should be used. In general, we need to know when RCT explanations are appropriate: that is, we need to know when we should use them. Further, we need to know when RCT explanations are sufficient by themselves. Sometimes, they will only be part of a broader explanation, but not yield the whole explanation by themselves.

Our knowledge of preferences presents a limit on when they are appropriate. An important precondition on the suitability of such a model is that we have reasons to believe the agents have the preferences, and beliefs that the model says they do. Often these reasons will be based on a rough and ready view of human psychology. However, if the only evidence of preferences is the behaviour itself, then it is entirely circular to invoke

these preferences to explain the behaviour. Thus, we can state one necessary condition for an RCT explanation to be appropriate:

1. RCT explanations are appropriate only if we can find out what the agent's preferences, beliefs and decision-procedure is.

Earlier, we noted another limit, which was that subconscious action can't be explained.

2. RCT explanations are appropriate only if the agents act on conscious motivations.

I leave open the question of whether there are further constraints on the appropriateness of RCT explanations, and so I make no claims to these conditions being jointly sufficient.

With regard to the sufficiency of RCT explanations, the all important question is whether these beliefs and preferences call out for their own explanation.

3. RCT explanations are sufficient by themselves just in case the RCT explanation is appropriate and the components of the model do not need explanation themselves.

If we feel these components need explanation themselves, then invoking them in an RCT explanation is not enough. For example, if we feel we need an explanation of why genders wear different clothes, then a model which posits preferences for the clothes associated with one's gender will not be complete by itself. On the other hand, if we don't feel a need to ask why people are moved by norms of fairness, then a model which posits preferences for behaving fairly can be a complete explanation by itself.

It is important not to confuse the charges of inappropriateness and insufficiency. If an RCT explanation is inappropriate it should be jettisoned. But if it is merely insufficient,

then it may well be prudent to retain it. In fact, it will often be a virtue of the RCT framework that it directs future research towards what is needed to make our explanation sufficient: it makes salient which component of the model is the one which needs further explanation.

I take it as a virtue of this account of appropriateness and sufficiency that it fits well with actual practice. In particular, it explains why RCT has been so successful at explaining the effect-oriented egoistic behaviour I have been at such pains to deny it is committed to. We feel very confident in our belief that, in competitive contexts like markets, agents' preferences will be egoistic and based on the effects they can get in interactions: they don't care about what actions they perform in themselves, but only the (financial) effects of doing so. Further, we don't feel that we need these preferences need their own explanation. Where RCT has been successful, the conditions for appropriateness and sufficiency have been met. (Of course, this does not show that market-based contexts are the only arenas where these conditions will be met).

Still, even when an RCT explanation is appropriate, and would be sufficient, it doesn't follow that it should be used. Our decision about whether to do so must be made in light of what alternative explanations are at hand, and one would need to make a comparative assessment of their respective theoretical virtues. The final decision will be based on the normal criteria by which theories are evaluated against each other: it will be an a posteriori question that turns on how fruitfully, simply and powerfully the theory explains the

evidence we have. Nothing I have said in this essay suggests that RCT will be able to do this better than alternatives. Rather, what I hope to have shown is that there is no a priori reason why RCT should be ruled out before it faces the tribunal of evidence: it is a mistake to assume that the framework is fundamentally incapable of explaining norm-driven behaviour, as some people assume.

Bibliography

Anderson, Elizabeth 2000, "Beyond Homo Economicus: New Developments in Theories of Social Norms" *Philosophy and Public Affairs*, Vol. 29, No. 2. (Spring, 2000), pp. 170 – 200

Davidson, Donald, 1963. "Actions, Reasons and Causes" reprinted in (Davidson 1980)

- 1971. "Agency". Reprinted in (Davidson 1980).

- 1980. *Essays on Actions and Events*. Oxford University Press: Oxford.

Elster, Jon 1985 "The Nature and Scope of Rational-Choice Explanation" reprinted in MM.

Hollis, M. 1977 *Models of Man*, Cambridge: Cambridge University Press.