Collective Action and Rational Choice Explanations

Randall Harp
University of Vermont
70 South Williams St, Burlington, VT 05401
randall.harp@uvm.edu

ABSTRACT

In order for traditional rational choice theory (RCT) to explain the production of collective action, it must be able to distinguish between two behaviorally identical possibilities: one, that all of the agents in a group are each performing behaviors in pursuit of a set of individual actions; and two, that all of those agents are performing those behaviors in pursuit of a collective action. I argue that RCT does not have the resources necessary to distinguish between these two possibilities. RCT could distinguish between these possibilities if it were able to account for commitments. I argue that successful rational choice explanations of collective action appeal to commitments, and distinguish this way of explaining collective action from a general class of explanations called plural subject (or team reasoning) theories.

I. INTRODUCTION

One virtue of a good explanatory model is that it is capable of explaining why event e occurs rather than event e', for some domain of events E which is within the explanatory purview of the model. For certain domains of events, rational choice theory (RCT) is a good explanatory model in this respect. Given a set of actions A, where every $a \in A$ is an *individual* action, RCT can be an important part of a good explanation for why agent X performed action a_1 rather than action a_2: namely, because X is rational, and because X judges that a_1 is more likely to produce a more highly-valued outcome than action a_2.

As I argue in this paper, however, for another domain of events RCT is not a very good explanatory model. Suppose that b_i is a group of individuals walking *individually* along a path towards a common location (for example the summit of a mountain), while b_c is a group of individuals walking *together*, along the same path, towards the same summit. These seem to be two different events, even though the behaviors of the individuals in question are (by stipulation) the same across b_c and b_i. In this case, I will argue, RCT cannot explain why, given the occurrence of the constituent behaviors, it is event b_i that occurs rather than b_c or vice versa. The problem is a general one: if B is a set of actions, and actions $b \in B$ consist of both *individual* and *collective* actions, then RCT is not a good explanatory model over this set of events because, given particular actions b_i and b_c, where b_i is an individual action, b_c is a collective action, and b_i and b_c are *behaviorally indistinct*, RCT cannot explain why b_i occurred rather than action b_c.[1]

If traditional RCT faces problems in distinguishing between individual actions and collective actions, what are our options? One option would be to eschew rational choice explanations altogether in favor of an explanatory strategy that makes no appeal to RCT at all. Many theorists working on collective action in fact take this route, as they either do not explicitly articulate a role for rational choice explanations, or they take no steps towards showing how their theory might be (at least partially) embedded in a rational choice framework. Of course, simply failing to appeal to rational choice theory in the course of giving an explanation of collective action is not itself a sign that one has judged the value of rational choice explanations and found them wanting; it might simply indicate a failure to consider the position. In any event, any theory that fails to appeal to rational choice explanations in distinguishing between individual and collective actions falls into this first option.

In this paper, I will not consider this option. Rational choice explanations provide useful resources for explaining actions in general, and if rational choice explanations can be marshaled to explain the differences between individual actions and collective actions, all the better. I will thus work under the assumption that we can develop a rational choice theory that contributes to an explanation of individual and collective actions, even if our rational choice theory needs to be modified from its standard form.

If we reject the first option, and thus are committed to some kind of rational choice explanation, there are two options remaining: we can either make relatively conservative changes to the basic framework of rational choice explanations, so that the rational choice theory that results still involves all of the resources of rational choice theory, though those resources might be changed or employed in different ways; or we can make more significant changes to the basic framework of rational choice theory, giving it new resources entirely. I will argue in this paper that conservative changes to traditional rational choice theory will not be

sufficient, and that we will need to adopt the third option. If we want to use rational choice theory in order to explain the differences between individual action and collective action, we will have to change our rational choice theory in significant ways in order to do it (though, I argue, what is left over will still be recognizable as a rational choice theory).

My argument proceeds as follows. First, I describe the resources that a typical RCT makes available to us for explanation; I call this RCT *traditional RCT*. I then argue that traditional RCT cannot distinguish between individual actions and collective actions. I survey the available conservative modifications of traditional RCT that can account for collective actions and I conclude that none are likely to be successful. Of the more significant modifications of traditional RCT, there are two viable candidates: what I will call *plural subject theory* and what I will call *rational choice with commitments*. I finally argue that plural subject theory faces problems that rational choice with commitments does not, and that the best candidate for a workable version of RCT that explains collective action is rational choice with commitments.

II. RATIONAL CHOICE EXPLANATIONS

Rational choice explanations can be important components of good explanations. In particular, rational choice explanations are useful as part of an explanation of *intentional action* because rational choice explanations effectively serve to highlight the agent's *reasons* for acting. In this section I will give an account of traditional rational choice theory and rational choice explanations, and I will describe the resources that traditional rational choice theory uses to provide explanations.

Note that I am not claiming that rational choice theory is *sufficient* for a complete explanation or account of action, nor am I suggesting that anyone ever argued for such a claim. A more complete account of action will require a discussion of: what an agent is; of the difference between actions and mere behaviors (and thus of the connection between reasons and actions); and of what it means for an agent to perform an action *for those reasons*. If we want to give a full etiological explanation of an action, we will need to specify those psychological and physiological facts about agents that enable them to perform the behavioral components of actions. Likewise, we might want to account for the difference between the three following possibilities: being *caused* to perform some behavior; *having* a (normative) reason to perform some behavior; and *endorsing* a (normative) reason to act. We might want to distinguish between different types of (normative) reasons that we have to act, as for example between reason-providing desires, reason-providing intentions, values, reason-providing beliefs of normative properties, etc. No rational choice theory will be able to provide answers to these questions.

Nevertheless, rational choice theory can, I argue, be an important *element* in our more complete explanation or account of action. Consider the following condition:

> **Optimality**: Some behavior B, performed by agent X, is not an intentional action unless: (i) X is in possession of some reason R that recommends A; and (ii) X is not in possession of some reason R' that recommends some non-compossible action B, where R' is stronger than R.

Optimality is somewhat controversial, in part because it seems to deny the possibility of actions that are both intentional and akratic.[2] Nevertheless, I think that we ought to accept Optimality, or something very like it, as an important necessary condition on intentional action. What Optimality points to is the close connection between intentional actions and reasons—that when we act intentionally, we necessarily act in light of what we take ourselves to have reason to do. To the extent that we judge our behavior to be misaligned with our judgment of reasons, we also find ourselves alienated from our actions and in doubt of our agency.[3]

If we accept Optimality, then any explanation of intentional action must account for Optimality's being satisfied. This is the primary benefit of employing rational choice explanations in explanations of action: a rational choice explanation entails that Optimality is satisfied, and accounts for why Optimality is satisfied. We do not need to believe that rational choice explanations are the *only* way to account for Optimality—only that they are a good way. If so, then rational choice explanations can be an element of a good explanation.

There are many different things that we might mean by 'rational choice explanation'; let us, then, stipulate the following interpretation for the present discussion. A rational choice explanation is one which explains agent X's *performing* some action a (or *choosing* to perform some action a; we can take the two to be the same thing) in virtue of the fact that X is rational, and a is judged by X to be an *optimal* action given the decision problem G that X sees herself as facing. Note that, on this account of a rational choice explanation, the reasons that explain the action are *subjective*—that is, the explanation goes by way of the agent's *beliefs* about the decision problem that she faces, and about the desirability of her available actions, rather than by way of the *objective facts* about the decision problem she faces or about the desirability of her actions. This ensures that our rational choice explanation does not bypass the agent's agency. In order for some behavior to be an action, the performing agent must stand in the right intentional relationship to the behavior, and rational choice explanations cannot appeal to these intentional properties if rational choice explanations appeal strictly to objective features of decision problems. All rational choice explanations are, at bottom, *maximizing* explanations.[4] The behavior itself is identified as maximizing some value function, and the agent's rationality consists in the agent's choosing the optimal outcome.[5]

I will use the term 'rational choice theory' to cover any formalization of a rational choice explanation. There are different ways that this formalization might be done; I will use the term 'traditional rational choice theory' to represent that class of rational choice theories which can be traced back to Savage (1954) and von Neumann and Morgenstern (1944), and the abbreviation 'RCT' will henceforth refer to traditional rational choice theory. Following Sugden (1991), I am interested only in the normative interpretation of RCT, in which the model is to be taken as a guide to decision making, rather than a descriptive interpretation of RCT in which the model is taken to predict an agent's actual behavior. RCT is a form of rational choice explanation because RCT is based upon expected utility theory: given a particular utility function, and given a particular decision problem, RCT entails that a rational agent will choose that behavior which maximizes that utility function.

As a generalization of rational choice theory, we model the decision problem G (for agent X) as follows: G consists of a set of *agents*; a set of *behaviors* for each agent;[6] a set of *outcomes* which can be produced; a probabilistic *outcome function* which maps behavior profiles onto outcomes; and a *value ordering* for each agent over outcomes, such that if an

agent judges herself or himself to have more reason to bring about outcome o_i than o_j, then the agent's value function will rank o_i higher than o_j; and a derivative value ordering over behaviors (so that if behavior b_i brings about more preferred outcomes than b_j, then the agent will more highly rank b_i than b_j).[7]

In addition, RCT uses a *representation theorem* to impose additional rational constraints on the values that the above features can take: so long as an agent's value ordering satisfies certain constraints, then that value ordering can be represented by a unique utility function and probability function, such that we can calculate an expected utility for behavior in G. There is much to be said about the status of representation theorems (see e.g. Meachem and Weisberg 2011), but because my primary interests in this paper lie elsewhere, we can follow Savage in our choice of representation theorem here.

Let us say that, for some decision problem G, an action a^* is considered *rationalizable* for agent X so long as a^* contributes to the production of the most highly-valued outcome it is possible to produce given X's value-ranking of outcomes. In *parametric* choice situations—that is, when the number of agents is one—the rationalizable actions are those which straightforwardly maximize the value of the outcome produced by the outcome function.[9] In situations of *strategic* choice—when the number of agents is two or more—some behavior is rationalizable for agent X when the behavior is a part of an action profile which is a *solution* to the decision problem. We can adopt Nash equilibria as a solution concept below, without loss of generality.

It is now apparent why rational choice explanations in general, and RCT specifically, can be useful for explanations of fully intentional actions. Suppose X is rational, and X performs some action a^* (where a^* is an intentional action and not mere behavior). If Optimality is to be satisfied, X must judge a^* to be better supported by her *reasons* at the time of the action. And we can nicely capture the idea of one action being more rationally supported by an agent's reasons than another by appealing to a decision problem in which the behaviors underlying the actions are represented, and of which the agent judges that one behavior is more likely to produce an outcome more in accordance with the agent's reasons for acting than the other behavior. Any explanation of action qua action must appeal to reasons somehow, and any explanation of a choice of one behavior over another must appeal to Optimality. Rational choice explanations, and RCT, seem to provide these.

III. THE INADEQUACIES OF RATIONAL CHOICE THEORY FOR COLLECTIVE ACTION

I am going to argue that RCT as currently constituted does not have the resources to provide proper explanations of collective action. Let us now make this objection precise. My objection against RCT is that it fails to be able to resolve what I will call the *discrimination problem*. The discrimination problem can be stated as follows: given a set of behaviors performed by a set of agents, and given a set of justifications for those behaviors (in the form of value functions), we do not yet have enough information to determine whether that set of behaviors jointly constitute a collective action or a set of individual actions. If the discrimination problem holds, then RCT will be of limited use in explaining collective actions (as opposed to the production of a set of behaviors), because RCT does not have the resources to explain why that set of behaviors is a collective action and not an individual action.

We should draw a distinction between the discrimination problem and another objection which is often directed towards RCT, which we can call the *outcome inadequacy problem*. According to outcome inadequacy, we can, independently of our theoretical commitments to any particular version of RCT, identify certain outcomes or states of affairs as *cooperative outcomes*. Once we identify certain outcomes as cooperative outcomes, then a putative challenge to RCT arises: is the act profile that leads to that outcome part of the solution set within RCT? If it is not, then that seems to be at least one mark against RCT: RCT does not entail that it is rational for agents to cooperate. As an example, the challenge that the Prisoners' Dilemma (PD) poses for RCT is an outcome inadequacy challenge: our intuitions are that the mutual cooperation outcome in a one-shot PD game is the 'rational' outcome for the agents to bring about, and yet mutual cooperation is not a solution to the game. The 'Hi-Lo' game, as discussed by Sugden (2000) and Bacharach (2006), is another example of an outcome inadequacy challenge to RCT. Though outcome inadequacy arguments are important challenges to RCT and have been much discussed, my objection to RCT is not an outcome inadequacy challenge.

Setting aside the outcome inadequacy worries, one might worry that even if the discrimination problem holds, that still does not reveal there to be any problems with RCT. After all, RCT is not sufficient for explaining action. Why should we not just keep RCT as it is, and resolve the discrimination problem using some other resources?

Note first that the challenge presented against the use of RCT in solving the discrimination problem is *not* a challenge against RCT itself. RCT can be a perfectly useful explanatory tool in one domain, and a less useful tool in another. The challenge to be presented here, then, is a worry about the use of RCT in one particular domain: namely, in explaining collective action specifically. Explaining collective action requires that we solve the discrimination problem somehow, and it is plausible to think that RCT, or some variant of it, will allow us to solve the discrimination problem. If RCT cannot solve the discrimination problem, I argue, then any account of collective action which tries to use RCT in that way will be inadequate.

What reason do we have to think that the discrimination problem holds for RCT? A more full argument would require us to fully canvas all of the resources that RCT has available to solve the problem; this will be done more carefully in section IV. Here, it will be useful to establish prima facie reasons why one might think RCT vulnerable to the discrimination problem.

The discrimination problem is generated by two premises which relate the components of individual action with the components of collective action. The first premise we can call *behavior compositionality*. According to behavior compositionality, the *behavior* performed by a group engaged in a collective action is composed without remainder of the behaviors of the individual agents who constitute the group—i.e. there is nothing left over when you subtract the behaviors of the individual agents from the behavior of the group performing a collective action.[10] The second principle we can call *act variability*, which holds that for any behavior α, an agent can *justify* the performance of α as either contributing to the production of an individual action, or as contributing to the production of a collective action.[11]

Keeping in mind the distinction between behavior and action, behavior compositionality is, I think, uncontroversial. In order for behavior compositionality to be true, collective actions

must have some behavioral basis, and the behaviors that constitute that basis must be performed by the agents who make up the collective—there is no additional behavior that is performed by the collective itself but not by some individual member of the collective. To my knowledge, this claim is not denied even by those who advocate for the most holistic or non-reductive theories of collective action.

Act variability is also widely supported; we can see versions of it in e.g. Gilbert (1989) and Searle (1990). Once we allow that the same behaviors might constitute either a collective action or an individual action, we face the question of what the difference between the two consists in. One partial answer is that each individual agent sees the behavior as justified either because it contributes to an individual action or to a collective action. Consider Searle's example of people running in from the rain: Searle suggests that a common set of behavior, namely a set of individuals running to a central shelter in a specific fashion, can either constitute a set of individual actions (as when all of the participants just want not to get wet from the rain) or a collective action (as when all of the participants view their behavior as elements of a complex choreographed dance). In both cases, the exact same behavior is performed. However, the participants justify the behavior differently depending on whether they are performing an individual action or their part in a collective action; this is what is meant by act variability.

To consider one more example from the literature, consider Gilbert's example of two people walking together (Gilbert 1996). The same set of behaviors—walking a certain route at a certain speed—might either constitute a set of individual actions or a collective action. Likewise, the agents themselves might understand their behavior as justified either in virtue of its contributing to an individual action, or in virtue of its contributing to a collective action.

I think that act variability, like behavior compositionality, ought to be fairly uncontroversial; it requires us to accept that collective actions consist of behaviors performed by individuals, and that the content of the behaviors themselves do not distinguish between being a part of an individual action and being a part of a collective action. Rejecting act variability would require us either to deny that the behavior can ever be identical across individual and collective action cases, or to deny that individuals can justify their own contributions to collective action. To deny the latter claim is to deny that individual reasoning can contribute to instances of collective action, which is implausible. Denying the former claim is equally implausible, as it suggests that the behavior itself changes depending on whether it is a part of a collective action or an individual action. If, however, we accept Searle's claim that agents might be fundamentally mistaken about whether they are indeed contributing to a collective action—they might, for instance, be confused as to whether the other participants in the collective action are actual agents or cleverly designed automata which causally affect the world without possessing agency—and if we also accept the claim that agents know the behavior they perform, then we cannot deny that behavior might be identical across individual and collective action cases.

Behavior compositionality and act variability pose a prima facie challenge to RCT because RCT does not seem to have the resources necessary to distinguish between the individual action case and the collective action case. Let B be a set of behaviors $\{b_1, \dots, b_n\}$, each behavior corresponding to each of the n agents in some decision problem. According to behavior compositionality, there are two sets of actions, c_b and i_b, which have B as their

behavioral basis—that is, in both c_b and i_b every agent $j \in n$ performs their designated behavior b_j. The set c_b consists of one collective action with the behavioral basis B; c_b is the collective action that occurs when every agent performs their behavioral part. The set i_b consists of n individual actions, each of which has as its behavioral basis the corresponding behavior b_j; i_b is a set of individual actions, though each agent is performing the same behavior as they would have in c_b.[12] (To use a hiking example, c_b is a collective hike, where each member walks a certain speed along a certain path, while i_b is a set of individual hikes, but where each member walks at the same speed and along the same path as in c_b. According to act variability, in order for RCT to distinguish between c_b and i_b, RCT needs the resources to determine whether, for each agent $j \in n$, the behavior b_j is justified as contributing to c_b or i_b. Justification in RCT is captured by value functions. However, it is a necessary feature of RCT that each agent only has one value function, and since that value function ranks behaviors, the same behavior will be ranked identically whether it is part of c_b or i_b. If the preceding is correct, RCT does not have the resources needed to account for act variability, and so the discrimination problem is a real problem for RCT.

IV. WHY TRADITIONAL RATIONAL CHOICE THEORY CANNOT BE REHABILITATED

We saw in section III that RCT cannot resolve the discrimination problem by appealing to different behavioral profiles of c_b and i_b because the behavioral profiles are the same. We also saw that RCT cannot resolve the discrimination problem by appealing to agent j's different justifications for $b_j \in c_b$ and $b_j \in i_b$, because j only has one value function and that value function must evaluate the two behaviors identically. Thus, if RCT is to be able to resolve the discrimination problem, it must do so in some other way. As I argue in this section, however, RCT cannot resolve the discrimination problem in any other way. If we look at all of the resources that RCT has, we can see that none of them are rich enough to enable us to distinguish between c_b and i_b. If we can only appeal to the standard elements of RCT, then any attempt to solve the discrimination problem will always run afoul of either the behavior compositionality condition or the act variability condition.

As a reminder, the structural elements of RCT as presented in section II consist of a set of agents, a set of behaviors for each agent, a set of outcomes, an outcome function, and a value ordering over outcomes which entails a value ordering over behaviors for each agent. If RCT is to be able to resolve the discrimination problem with nothing more than minimal revisions, then it must be able to find a way to do so employing the resources described. We have already seen why appealing to the set of behaviors itself is unlikely to be of much help; the discrimination problem presupposes that the chosen behaviors for c_b and i_b are the same. Likewise, appealing to the value ordering is unlikely to be helpful; since each agent has only one value ordering, that value ordering must evaluate each agent's behavioral contribution to c_b in the same way as it evaluates that agent's behavioral contribution to i_b.

What other options do we have? Appealing to the set of agents will not be helpful without making more radical changes to RCT. If the set of individual actions i_b can be explained in part by the fact that the behavioral profile B which underlies i_b is chosen by the n agents in the decision problem, then it is not clear why appealing to a different set of agents n' could explain why that same behavioral profile B brought about the collective action c_b instead. (The

plural subject model of rational choice can be seen as a radical way of modifying the set of agents; we will look at that approach more fully in section VII below.)

Another intuitively plausible approach is to try to account for the difference between c_b and i_b by appealing to differences in the outcomes available in the decision problem. According to this approach, outcomes correspond to actions, and since individual actions and collective actions are different, there must be different outcomes that correspond to them. This approach, though plausible, fails to save RCT. Suppose there are two outcomes that correspond to the different actions that might be produced: one outcome o_c corresponds to the collective action c_b, and the other outcome o_i corresponds to the set of individual actions i_b. Those outcomes must be produced by the decision problem's outcome function, and that function takes behavior profiles as inputs and returns outcomes as outputs. Since (according to behavior compositionality) c_b and i_b have the same behavior profiles, the outcome function must map them onto the same outcome (or onto the same outcomes with the same frequency, if the outcome function is stochastic). Even if we want to allow that a collective action is a different outcome from an individual action, we cannot do so just by adding new outcomes into the decision problem.

The preceding strategies all try to solve the discrimination problem by modifying the existing elements of RCT, in particular by trying to broaden the scope of those elements so that they can accommodate collective actions as well as individual actions. As we saw, these strategies are not successful because they all entail a violation of behavior compositionality. Rather than attempting to account for collective action by modifying existing elements of RCT, then, another possibility is that collective action can be explained through *structural* relationships that hold between the attitudes of the agents in the decision problem; this approach, as we will see, fails the act variability condition.

Accounting for collective action through structural relationships means appealing to the way that one agent's attitudes are related to another. For example, one might think that the difference between c_b and i_b lies primarily in whether the agents' value functions are *identical* over outcomes. If the value functions are identical, then the agents are cooperating when they choose the behaviors that constitute B, and thus they are bringing about c_b; if the value functions are not identical, however, then they are not cooperating, and they are bringing about i_b. It's not clear to me whether anyone actually argues for this strategy, at least not as described; Bacharch (2006) comes closest, though in addition to uniformity of value functions he also requires that agents engage in a distinctive form of reasoning called team reasoning, which we will look at in section VII below. It is worth considering the strategy, however, as an attempt to capture the intuition that cooperation occurs when agents are in agreement not only about what needs to be done, but (roughly) share in their judgment of reasons *why* it needs to be done. Admittedly, RCT cannot easily account for agents sharing reasons why some outcome ought to be brought about; value functions are too crude to properly account for *reasons why*, and can only really account for comparative judgments of *reason to*. However, a plausible necessary condition for two agents sharing reasons why something ought to be done is that the agents agree in their comparative judgments about which of a set of actions in a decision problem agents have reason to do. (I will not argue for this as a necessary condition, and nothing in my argument depends on this claim being true; readers who are skeptical can skip this discussion with no consequences for my overall argument.)

The problem with this view is that it's not clear why sharing a value ordering would help at all to explain why a chosen behavioral profile is cooperative action. Consider the Hi-Lo game in Figure 1 which we can call Quick-Slow. In Quick-Slow, both agents prefer to walk together than separately, and they also each prefer to walk quickly together than to walk slowly together. Of the two non-coordinative outcomes, both X and Y are indifferent: they do not care whether they walk quickly and the other agent walks slowly or vice-versa. X's value function is thus identical to Y's value function, so if the difference between c_b and i_b is whether the agents share a value function, then we should conclude in this case that X and Y are cooperating when they both choose to walk quickly. (We suppose that the (*quick*, *quick*) equilibrium is sufficiently salient to be judged the unique solution to the game by both agents.)

		Y: *slow*	Y: *quick*
X	*slow*	$2, 2$	$1, 1$
	quick	$1, 1$	$4, 4$

Figure 1: Quick-Slow

However, this is too hasty; the fact that X and Y share *individual* judgments about the comparative desirability of the four outcomes does not give us any reason to think that the chosen behavioral profile is cooperative. X and Y might each only have individual reasons for preferring to coordinate rather than not coordinate: perhaps X is not entirely confident that she knows where she is going and has reason to believe that Y is going to the same place, while Y is fearful for his safety and prefers to walk in the company of others. Even if we allow that these respective motivations are common knowledge, we are not yet justified in claiming that X and Y's choosing to walk at the same quick speed is a *collective action*. Act variability tells us that each agent should be able to justify their performance of their behavior as either contributing to a collective action, or contributing to an individual action. But merely sharing a value function with someone else is not sufficient to make the difference between justifying an individual action and justifying a collective action; agents can share a value function even if they don't know they are sharing it, or even if they don't want to share it. Moreover, agents can share a value function even though they have no intention of contributing to a collective action. Consider the passengers who board the subway car and distribute themselves evenly throughout the car, knowing that everyone prefers an outcome in which they all have a maximum amount of personal space—the fact that they all desire this outcome and act to bring about this outcome does not imply that the subway riders are acting *collectively*. Coordinating on a preferred action profile is necessary for two agents to act collectively, but it is certainly not sufficient. Likewise, sharing a value function is certainly not sufficient for engaging in a collective action. (Indeed, I am going to argue below that unlike coordinating on a preferred action, sharing a value function is not even necessary.)

One possibility that we do *not* need to consider in detail here is the possibility that the difference between c_b and i_b lies in whether the agents' value functions recommend *specific*

outcomes. To return again to the Quick-Slow example above, we might think that the agents are engaged in a collective action so long as their value functions recommend the coordinative outcomes (where they both walk quickly or they both walk slowly) over the non-coordinative outcomes. We do not need to consider this strategy because it depends on our pre-theoretical identification of the coordinative outcomes as cooperative outcomes. This is the same assumption that is made by outcome inadequacy theorists, and this argument is essentially a version of the outcome inadequacy argument against RCT. As such, considering it in detail is outside the scope of this paper.

V. COMMITMENT MODELS OF RATIONAL CHOICE

In the previous section, we looked at whether it was possible for RCT to answer the discrimination problem while making only conservative changes. Changes to RCT are 'conservative' so long as they continue to use the same resources that RCT has available for ordinary rational choice explanations, even if changes need to be made to those resources in order to account for why the behavioral profile B might make up c_b in one instance but i_b in another instance. As we saw, it is not possible to explain the difference between c_b and i_b through only conservative changes to RCT.

If we are to resolve the discrimination problem through a rational choice explanation, then, we are going to have to make major changes to RCT in order to do so. There are two plausible options in the rational choice literature for doing so: *plural subject* models of rational choice, and *commitment* models of rational choice. There are examples of both kinds of modifications of RCT in the literature, though my intention here is to talk about the two kinds at a fairly general level. I will thus appeal to specific details about specific models of rational choice when appropriate, but my interest in doing so is to provide examples for how the modifications would work, and not to talk about these models specifically. (In fact, though I will argue that commitment models of rational choice are best suited to providing rational choice explanations of collective action, I also think that none of the current leading models are ideally suited to accounting for collective action—though I will not argue for that claim in this paper.)

I want to define what I mean by 'commitment'; the term will be used prescriptively here, though it will be clearer below why I use this term when I connect commitments with intentions. Elster (2000) has written extensively on commitments (he uses the term 'essential constraints'), and while the account that I give here is similar in some respects, my discussion of commitments is more in the context of rational choice explanations than Elster's. Likewise, Sen (1985, 2005) has written on commitments and rational choice, and my treatment of commitments is informed by his work as well. Commitments are exogenously specified elements of a decision problem that determine the rationality of one's choices. The term 'commitments' here will be used primarily as a contrast with other rational requirements that agents face in decision problems—in particular, those generated by agents' value functions over outcomes. Let us recall how this works: on the basis of an agent's value function and the representation theorem, we can construct a cardinal utility function for each agent which represents how much utility an agent derives from each outcome in the decision problem. It is in virtue of this utility function that we can say that some outcomes are *better* or *worse* for the agent to bring about, and it is in virtue of this ranking of outcomes that agents' behaviors are judged more or less rational to choose.[13] It might be that a fully rational agent is one who

always brings about the outcome which produces the maximum amount of utility possible—this is the maximizing conception of rationality which rational choice explanations assume—but we can still always assess *how far* some outcome is from being optimally rational by comparing the cardinal utility of the outcome with that of the other outcomes in the decision problem.

In contrast, the sorts of rational requirements generated by commitments do not admit of degrees; rather, they are all-or-nothing requirements. When one possesses a commitment, one has a rational obligation to do that which the commitment requires. Failing to satisfy one's commitment is thus simply failing to satisfy one's rational obligation. If I commit myself to eating fifty eggs, then nothing that I do short of eating fifty eggs will satisfy my commitment. Moreover, there is no way to *better* satisfy one's commitment qua commitment. If my commitment is to eat fifty eggs, then eating fifty eggs at one time and later in the day eating another fifty eggs does not satisfy my commitment more fully or completely than if I just ate fifty eggs. Now, if my commitment were to eat *only* fifty eggs, then of course eating one hundred eggs would violate my commitment. However, if my commitment is to eat *at least* fifty eggs, then eating fifty-one eggs is no better or worse a way of satisfying my commitment than eating fifty eggs.

Satisfying commitments, then, is strictly binary: for any given commitment either one has satisfied it, or one hasn't, and performing some behavior either will satisfy the commitment, or it won't. There can be no cardinal ranking of utilities with respect to the satisfaction of commitments. In this way, commitments order outcomes differently than value or utility functions, which admit of degrees. This is not to say, however, that there is no possible way to rank various outcomes which all satisfy one's commitment—only that the commitment itself will not do the ranking. If my commitment is to eat at least fifty eggs, then the outcome in which I eat fifty eggs is of the same status as the outcome in which I eat a hundred eggs, at least in the eyes of the commitment itself. I might, however, have other reasons to prefer the outcome in which I eat only fifty eggs: propriety, health, common sense, and so forth.

Further, there is some flexibility in how we describe our commitments. My commitment to eat at least fifty eggs is satisfied only in those outcomes in which I have eaten at least fifty eggs. My commitment to eat as many eggs as is physically possible is satisfied only in those outcomes in which I have eaten as many eggs as I can physically manage. In the latter case, it might seem as though eating the fifty-first egg is 'better' at satisfying the commitment than merely eating the fiftieth egg. This is not right, however. If my commitment is to eating as many eggs as possible, then the fifty-first egg is not a better way of satisfying the commitment—it is the *only* way of satisfying the commitment. (And so on for the fifty-second egg, and the fifty-third egg, etc.) Commitments can be satisfied or not; value functions can be maximized or not. Commitments and value functions both impose rational constraints upon an agent's deliberation; rationality requires that commitments be satisfied and value functions be maximized. However, commitments are different from value functions in how the non-rational outcomes are treated: commitments treat them all the same, while value functions treat the non-optimal outcomes differently.

We can distinguish between three types of rationality: *value rationality*, *minimal commitment rationality* and *maximal commitment rationality*. An agent is value rational if and only if the agent resolves a decision problem by bringing about that outcome which is the most highly valued outcome possible given the beliefs and value functions of the other agents in the

decision problem. An agent is minimally commitment rational if and only if the agent resolves a decision problem by bringing about *some* outcome which is compatible with the agent's commitments—that is, if the agent brings about an outcome which is not proscribed by the agent's commitments. Maximal commitment rationality is a combination of the preceding two conceptions: an agent is maximally commitment rational if and only if the agent resolves a decision problem by bringing about the most highly valued outcome possible which is also compatible with the agent's commitments. Commitments and value functions are thus theoretically independent; commitments do not logically take lexical priority over value functions, or vice versa. The agent who is commitment rational (either minimally or maximally) prioritizes commitments over value functions, and the agent who is value rational prioritizes value functions over commitments.

As mentioned above, this use of the term 'commitment' is stipulative. Commitments are constraints on rational decision-making, and they impose different constraints than those imposed by value functions. Moreover, the two kinds of constraints are not in conflict; we can model both commitments and value orderings in the same model of rational choice. Doing so leads to three questions: first, why would we want to do so? Second, are there are any models in the literature that do attempt to capture commitments as I have described them? And third, supposing we were to model rational choice using both commitments and value functions, should we consider the resulting model a rational choice model? I want to answer these three questions in the remainder of this section.

Taking the first question first: supposing that the view of commitments that I presented above is coherent, why would we want to include them in a model of rational choice? Commitments, in my view, are exogenously specified constraints on practical deliberation that render any choice contrary to that commitment to be ipso facto irrational. If I have a commitment to eat fifty eggs, and I face a decision problem in which one choice results in my eating fifty eggs while another choice results in my not eating at least fifty eggs, my commitment would entail that the latter choice is irrational and the former choice is rational, at least as far as the commitment is concerned. There might be other factors that render the former choice irrational, but if a choice is irrational according to some commitment, then that choice cannot become rational again unless that commitment is removed. This view about how commitments work is very similar to how intentions are said to work in the literature on intentions.[14] According to this view, intentions generate rational pressure on agents to deliberate and act in accordance with those intentions. This rational pressure takes the form of a filter on deliberation: actions which are in accordance with one's intention are allowed to pass through the filter and be part of an agent's deliberation, while actions which are not in accordance with an agent's intention are rationally excluded from consideration. Intentions, in other words, are best modeled as commitments of the form that I have described.

We care about intentions because we care about the role that intentions play in explaining intentional action, both individual and collective. According to the received view, a necessary condition on agent X's acting intentionally is that X have an intention which entails the performed behavior. That is, if X's b-ing is to be an intentional action, then X must have an intention which entails the performance of b.[15] Likewise, in order for a set of agents to collectively perform some action intentionally, they must somehow share an intention which entails the performance of that behavior. More will be said about sharing intentions below. But

even before going into details, we have some reason to think that in order to provide a rational choice explanation of collective action, we need to employ a theory of rational choice which can account for intentions—that is, a theory of rational choice with commitments.

Once we are capable of modeling commitments or intentions in our rational choice theory, we then have resources to capture collective action that we did not have before. The main idea is this: agents are engaged in collective action so long as they act on the basis of their *shared commitment* to bring about some outcome. This is similar in outline to the dominant tradition in collective action, which holds that a necessary condition for agents to be engaged in a collective action is that they share an intention (Bratman 1993, Velleman 1997, Gilbert 2003, Roth 2004, Bardsley 2007, Alonso 2009, Tuomela 2005, Roy 2010). On my view, a necessary condition for agents to be engaged in collective action is that they share commitments with respect to the decision problem that they face. Thus if not every agent in a decision problem is committed to bringing about the same outcome or member of a set of outcomes, then the agents cannot be engaged in a collective action; any outcome they produce cannot contribute to c_b, but must instead contribute to i_b. If every agent shares a commitment, however, then the actions they perform can contribute to c_b.[16]

The second question posed above was whether there are any models in the literature that capture what I call commitments. While I will not develop a fully-realized model of rational choice with commitments here, we have every reason to believe that such a model can be developed. After all, several formal models capturing commitments have been developed, and while each model is different (and indeed, some are poorly suited for solving the discrimination problem), there are no technical challenges to developing such models. For example, the literature on 'Belief-Desire-Intention' models (or BDI models) tries to formally articulate the semantics of intentions, and models intentions similarly to what I call commitments. Prominent BDI models have been given by Bratman, Israel and Pollack (1988), Cohen and Levesque (1990), Rao and Georgeff (1991), and Woolridge (2000). Although BDI models are not ideally suited for rational choice explanations—they focus more on modeling the relationships that ought to hold between intentions, beliefs, desires, and related elements like goals, plans, choices and actions than they do on accounting for normative models of decision-making—they do suggest one promising avenue for formalizing commitments.

Besides the BDI literature, important work has been done in formalizing intentions in a model of rational choice. Van Hees and Roy (2008) and Roy (2009) model intentions in a way which capture the desiderata I have set forth for commitments: intentions for van Hees and Roy are filters of admissibility on rational decision making which serve to remove those behaviors from rational consideration which lead to outcomes incompatible with the intention. I will not discuss the models of van Hees and Roy in detail here; it suffices to note that they have provided one formalization of a rational choice theory with commitments. There are other possible formalizations; indeed, I make no claims as to whether or not the model presented by van Hees and Roy has all of the resources necessary to solve the discrimination problem. (For example, any acceptable formalization of commitments needs to have the resources necessary to talk about *sharing commitments* in an appropriate way.) But even if van Hees and Roy's models were not ideally suited to resolving the discrimination problem (and again I make no claims on this point either way), they would still give a foundation upon which any successful model could be easily built.

Returning to the third question posed above, we might also wonder whether any rational choice model with commitments remains a rational choice theory, or whether it becomes something else. Both Sen and Gauthier, for instance, have argued that rational choice theory needs to be able to accommodate something like commitments or constrained maximization; in doing so, however, they seem to be arguing that rational behavior need not be maximizing behavior (Gauthier 1986 and 1996; Sen 2005). It is thus useful to compare the claim that I made above—that rational choice models with commitments are still rational choice models because they are maximizing models—with Gauthier's and Sen's claim that commitments are useful precisely because they allow for rational agents that are not maximizing agents. Given the definitions we gave above about value rationality and commitment rationality, we can see that there is no conflict between Gauthier's and Sen's claims about commitments and my assertion that a rational choice theory with commitments is still a maximizing rational choice theory. According to rational choice theory with commitments, agents who exhibit maximal commitment rationality are rational agents. Agents with maximal commitment rationality *need not be value rational*, however. That is, agents with maximal commitment rationality can be rational even though they fail to bring about the outcome which maximizes their value function. In other words, it is possible to be a rational agent (of some kind) even if one is not value rational, and a rational choice model with commitments tracks that kind of rationality.

VI. THE ASYMMETRY BETWEEN INDIVIDUAL AND COLLECTIVE ACTIONS

Introducing commitments into our formal model of rational choice thus gives us resources for solving the discrimination problem that RCT does not possess. We might worry, however, that these resources make the problem bigger than it would otherwise seem to be. If we need to augment our rational choice model with commitments in order to explain collective action, the argument goes, why would we not also need to augment our rational choice model with commitments in order to explain individual actions? And if we did need commitments to explain individual actions, wouldn't that mean that RCT is completely explanatorily bankrupt? This conclusion would be a strong one—perhaps too strong. The conclusion itself could be avoided, however, if there existed some asymmetry between individual actions and collective actions, such that commitments might be needed to explain collective actions but not individual actions.

In this section, I want to argue that there *is* an important asymmetry between the individual action case and the collective action case, and it is in virtue of that asymmetry that RCT fails to solve the discrimination problem. Stated simply, the asymmetry is this: when a set of agents are performing a set of individual actions, the agents face rational pressure for their commitments to track their own individual value functions. When agents are performing a collective action, on the other hand, they face no such rational pressure; their commitments need not track their own individual value functions, but instead should track some value function of the group itself (where this is not necessarily reducible to the value functions of the individual agents).

Recall from the previous section that there are three useful notions of rationality that are in play when we consider a rational choice theory with commitments: value rationality,

minimal commitment rationality, and maximal commitment rationality. When agents are engaged in individual actions, value rationality and maximal commitment rationality ought to recommend that the same choices be made and the same outcomes be brought about. Another way to express this same idea is that for the agent engaged in an individual action, commitments are rational if and only if adopting the commitment allows the agent to maximize her or his own individual value function. The same does not hold in the case of collective action, however; agents engaged in collective action can rationally hold commitments even though those commitments do not maximize the agent's own individual value function.

Let us first see why rational agents engaged in individual action must hold commitments that maximize their individual value functions. Either an agent faces a parametric decision problem (in which either there is only one agent, or only one agent's choice is made on the basis of a rational consideration of the decision problem)[18] or the agent faces a strategic decision problem in which there is more than one agent. Consider first the case of a parametric decision problem. It has been noted by Gauthier (1986) that commitments do not rationally contribute to parametric decision problems. Suppose an agent X's value function recommends performing behavior b^*—that is, b^* has the highest expected utility of all actions under consideration. b^* is recommended because b^* leads directly to some outcome or outcomes, without the need for any strategic considerations of other agents' actions. Any commitment that X holds, then, will either proscribe b^* or not. If X's commitment does *not* proscribe b^*, then it is reasonable to wonder why X bothered forming the commitment in the first place—the commitment is not adding anything to the desirability of b^* that was not already present. If X's commitment *does* proscribe b^*, however, then we are left with the question of how this can possibly be rational. X is acting individually, and so X need not answer to anyone's value function except her own.[19] All of the considerations that X takes to be rational considerations for X's actions are included by definition in X's value function. Any commitment that proscribed b^* simply cannot be rational, because if it were rational it would have already been factored into X's value function.

The situation becomes somewhat more complicated when we consider strategic choice situations instead of parametric choice situations. The additional complication comes from the fact that agents cannot bring about outcomes on their own, but instead must rely on other agents; having a value ranking over outcomes is not sufficient in strategic choice situations to derive a value ranking over behaviors, as it is in parametric choice situations. For X and Y in a Prisoners' Dilemma (PD) game (Figure 2), each agent most prefers the outcome in which they confess (*defect*) but the other agent stays silent (*cooperate*). The second-most preferred outcome is that in which they both stay silent. Their third-most preferred outcome is that in which they both confess. The least preferred outcome is that in which they stay silent and the other agent confesses. Thus,

$$o_{\{dc\}} \succ_X o_{\{cc\}} \succ_X o_{\{dd\}} \succ_X o_{\{cd\}}$$

and

$$o_{\{cd\}} \succ_Y o_{\{cc\}} \succ_Y o_{\{dd\}} \succ_Y o_{\{dc\}}$$

where $a \succ_X b$ indicates that X strictly prefers a to b. Suppose we consider only X's value ranking over those four possible outcomes: do we have enough information to determine X's value function over the behaviors *defect* and *cooperate*? We do not; in order to determine X's

value function over behaviors we have to know both X's value ranking over outcomes and Y's value ranking over outcomes.

		Y: c	Y: d
X	c	o_{cc}	o_{cd}
	d	o_{dc}	o_{dd}

Figure 2: Prisoner's Dilemma

In spite of the complications that hold for strategic choice situations but not for parametric choice situations, the underlying moral remains the same: so long as agents are rational and performing individual actions, commitment rationality ought to coincide with value rationality. This is because for individual actions, intentions are rationality adopted when and only when they facilitate the production of the most highly-ranked outcome an agent is capable of bringing about. An agent's individual value function tracks that agent's reasons for acting, and individual actions must track individual reasons. Thus it is ultimately the agent's value function that determines to what extent some behavior is rational when an agent is performing an individual action.

This is not to say that there is never any reason to hold a commitment when performing an individual action. Commitments can transform games so that the chances of bringing about a favorable outcome increase for the agent holding the commitment. A clear example is in what is rather unfortunately known as Battle of the Sexes and which we will call Ballet or Soccer (BoS), where two agents each want to coordinate on an activity but they disagree about which activity they would like to coordinate on, ballet or soccer (Figure 3). In this example, X prefers that the two attend the soccer match, while Y prefers that the two attend the ballet. This game has two equilibria, and hence two solutions. If, however, X has a commitment to playing soccer (and Y knows about that commitment), that would reduce the number of equilibria in the game from two to one, forcing the two to coordinate on playing soccer. In this case, forming a commitment would be beneficial for X, as it would increase her ability to maximize her own individual value function.

		Y: *ballet*	Y: *soccer*
X	*ballet*	$1, 2$	$0, 0$
	soccer	$0, 0$	$2, 1$

Figure 3: Ballet or Soccer (BoS)

There is much more that needs to be said about commitments in actual games, of course. In the example described, X's commitment benefits her so long as Y does not have a corresponding commitment of his own to go to the ballet; if both agents hold a commitment, then they are doomed to fail to coordinate. We might worry, then, that either X and Y ought to both have a commitment, or neither can, and since it is clearly not value-maximizing for them to both hold a commitment, neither of them can hold one. We also might worry that commitments might just be *incredible threats* (also known as *noncredible threats*)—that is, a threat that act in a way that a rational agent would not actually do because doing so would be irrational. This is especially salient in games where choices are made asynchronously—if X had a commitment to go to the soccer game, but Y chose first and chose to go to the ballet, would it still be rational for X to choose in accordance with her commitment?

In fact, both of these concerns support my argument in this section, namely that in cases of individual action, commitments should promote the maximization of the agent's value function. If holding a commitment (rather than not) facilitates the maximization of an agent's value function, as in the case of the BoS game, then it is rational for the agent to hold the commitment. If holding the commitment does not facilitate the maximization of one's value function, then it is not rational to hold the commitment. As for incredible threats, they can be broken down into two claims: the first is an incredulity claim, and the second is a threat claim. The incredulity claim is that an agent would have no reason to act contrary to one's individual value function, regardless of the presence of a commitment. This claim is correct, I think: incredible threats are incredible precisely because they fail to maximize an agent's value function. An agent who is value rational will not hold a commitment which fails to maximize her or his value function. This should not, however, lead us to conclude that a commitment cannot actually be a commitment, or cannot constitute a genuine threat. By definition, an agent who is either minimally or maximally commitment rational will act in accordance with her or his commitments. We might wonder whether there *are* any agents who are commitment rational; we also might wonder whether being commitment rational is itself a rational thing. (We might, of course, also wonder this about value rationality.) We should not wonder, however, whether an agent who is maximally commitment rational would actually act in accordance with her or his commitment; the agent will by definition.

In the case of individual actions, value rationality and commitment rationality ought to coincide. This need not be the case for collective action, however. When agents are acting on behalf of collective actions, they can be rational by being maximally commitment rational *without* being value rational. This is what entails that explaining collective action requires a rational choice theory with commitments, but explaining individual action only requires RCT.

Why do collective actions not require that value rationality and commitment rationality coincide? Value functions reflect an individual agent's total judgment about the reasons that exist to bring about outcomes compared to one another. If collective action were to require that value rationality and commitment rationality coincide, as individual action does, then the implication would be that every rational collective action must maximize the value of the outcome produced according to the value functions of every individual agent. (By 'maximize' here I do not mean 'produce the most-valued outcome'; instead, I mean 'produce the most-valued outcome possible.')

However, collective action does not require that the outcome produced be judged as optimal according to the individual value functions of the agents that make up the collective. A group of agents acts collectively so long as each of the agents in the collective accepts the collective goal or the collective reasons for producing some outcome. But that collective goal, and those collective reasons, are not necessarily reducible to individual goals or reasons; they can be generated as values for the collective even if not all of the agents agree with this judgment. Indeed, they can be generated as values for the collective even if *no* agent agrees. (Agents *agree* with a collective reason when the outcome-ranking value function that corresponds with that reason is identical to the agent's own individual value function.) A family can rationally go on a vacation even when no individual member of the family wants that vacation—even if the vacation chosen is not a product of a process of bargaining (Sugden 2000). A gang begins to act when one member says "so, we're ready to go, right?" and the other members agree, even if no individual gang member judges the gang's action as optimal according to her or his value function (Gilbert 2003). Collective action requires the individual agents to subordinate their own judgments of the value of outcomes to the group goal or intention, and we can model the group's goal or intention as a commitment. Acting rationally on behalf of a group goal, then, requires that each individual agent be committed to acting on behalf of the group goal, even if doing so does not maximize their individual value function. Since they are not engaged in an individual action but rather a collective action, failing to maximize their individual value function is not irrational.

Thus, when an agent is doing her or his part to bring about the collective action c_b, it is no longer the case that maximal commitment rationality ought to coincide with value rationality. Value rationality tracks each agent's individual judgment about the desirability of c_b. In cases where agents successfully bring about c_b, however, the agents can disagree about the desirability of c_b and still contribute to producing c_b. In these cases, value rationality and maximal rationality will diverge. Being rational with respect to bringing about c_b does not require that one maximize individual value functions—it does not require value rationality—but it does require that agents pursue the common goal, which requires maximal commitment rationality.

VII. PLURAL SUBJECT MODELS VERSUS COMMITMENT MODELS

I argued in the previous section that there is an asymmetry between individual action and collective action, such that RCT (or other rational choice models without commitments) can explain individual actions but not collective actions. I have also argued above that a rational choice model with commitments can explain collective actions. It remains to be seen whether there are other viable rational choice models that can explain collective action. The strongest candidate is what I call *plural subject theory*; plural subject theory has advocates in both philosophical and formal rational choice accounts of collective action.[20] On the philosophical side, Gilbert is the most well-known advocate of plural subject theory, and has done the most to develop it (1996, 2001, and 2006). On the formal side, notable work has been done by Sugden (2000 and 2003), Gold and Sugden (2007), and Bacharach (1999, 2006).

The central idea behind plural subject theory is that the agent of a collective action is, properly speaking, a plural subject or team rather than a set of individuals. Different authors spell this idea out differently, but we can give a general account of plural subject theory here:

plural subject theory entails that collective actions are performed by a plural subject, or set of agents, on behalf of *collective attitudes*, where these collective attitudes are not reducible to individual attitudes; facts about the attitudes held by the individuals that make up the plural subject do not suffice to determine facts about collective attitudes. The collective attitudes play the same role in explaining collective action that individual attitudes play in explaining individual action. The relevant collective attitude here is the value function. Plural subject theory when embedded in a rational choice framework assigns one collective value function to the plural subject, and this collective value function controls what the team chooses and does: if some action profile b is considered optimal according to the collective value function, then the plural subject will perform the action profile b.

A collective value function by itself cannot explain why a plural subject performs some collective action c_b. Since c_b is made up of the set of behaviors b, and since the set b consists of n behaviors b_i for each $i \in N$ (where N is the set of agents who make up the plural subject), the plural subject theorist still has to account for why each agent $i \in N$ is motivated to perform her or his action b_i. According to the plural subject theorist, agents who act towards a collective action engage in *team-directed reasoning*. If an agent reasons according to team-directed reasoning, and if they are a member of a plural subject with some collective value function, then they infer that the rational action for them to perform is whatever their part is of the behavioral profile which is most highly favored by the collective value function. Thus team-directed reasoning gives agents a collective value function and also tells them what to do once they have it—namely, do their part. If every agent in some set N holds the same collective value function and also engages in team reasoning, it is easy to see that the plural subject will maximize the utility of the outcome produced (since the collective value function represents utility for the plural subject).

Plural subject theory thus explains collective action in the following way: when you have a set of agents, each of whom holds the same collective value function, and each of whom engages in team reasoning on that collective value function, then the agents will perform that behavioral profile b which maximizes the value of the collective value function. The outcome that is produced is c_b and not i_b because the outcome maximizes the collective value function rather than a set of individual value functions, and because the agents' choice was caused by their engaging in team reasoning rather than individual reasoning. In this way, plural subject theory solves the discrimination problem. Compare this to the approach that a rational choice with commitments model takes: according to rational choice with commitments, collective action occurs when each member of a set of agents has the same commitment with respect to which actions can and cannot be performed, and when each agent exhibits maximal commitment rationality, and when the behavioral profile b is judged rational according to each agent's individual value function after the commitment has filtered out incompatible outcomes. If the act profile b is produced in accordance with some shared commitment, then b supports c_b rather than i_b.

Which of these two rational choice-based explanations of collective action ought we prefer? The question cannot be fully resolved here, but I will give a few initial thoughts. First, it is worth mentioning again that *both* plural subject theory and rational choice with commitments represent a rejection of traditional RCT, and so either way my basic thesis (viz. that traditional RCT cannot solve the discrimination problem) holds. I do think, however, that

plural subject theory faces some prima facie challenges that rational choice with commitments does not face. First, plural subject theory requires that every agent share the same collective value function. Insofar as we identify the collective value function with a subjective recognition of reason to act, this suggests that, according to the plural subject theorist, collective action requires that every member of the collective agree about which outcome they most have reason to bring about. This is not consistent with a common understanding of collective action, however. Agreement about *reasons* is not requisite for collective action; agreement about *goals* is. Two agents who both grudgingly act to bring about some outcome are no less cooperative just because their contributions are contrary to their judgment about bestness.

Consider a football team which consists of a number of highly self-impressed egomaniacs. Each player on the team genuinely believes that the team's best chances of winning occur when the play is run through him. When considering the various outcomes (corresponding to which play to run), each football player ranks them differently ("we should run the play that features me"). Nevertheless, the team agrees to run a certain play. Are the members of the team performing a collective action? I think the answer is clearly yes. Do the members of the team share a collective value function? I think that here the answer is no. Their differing judgments about which play has the best chance of leading to a score are not *individual* value functions; they track what each player judges is best for the *team*. To see this, imagine that one player wants to be traded away from the team, and believes that will happen if no plays are called for him throughout the entire game. His individual value function accordingly favors no plays being called for him at all. Since he still thinks that he is the best player on the team, however, his value function from the standpoint of the team is that he ought to get the ball. ("Of course the team wants to get me the ball; that's the best chance we have of winning.") So the differing value judgments of all the players on the team are differing collective value judgments and not individual value judgments—and yet they can perform a collective action in spite of this disagreement. Plural subject theory has a difficult time accounting for this, but rational choice with commitments is able to solve this problem more easily: when the team agrees on a particular play, that play commits them in certain ways. Any disagreements they have in their individual value functions are now less important, because that disagreement is rendered irrelevant by the shared commitment.

The defender of plural subject theory might object that this captures the collective value function incorrectly. Couldn't the content of the collective value function be something like "value winning over losing"? And couldn't such a value function, if held by every member of the team, explain why the resulting behavior is a collective action rather than a set of individual actions?

The problem with this response is that the collective value function has to structure every outcome under deliberation by the members of the decision problem. If the team is deliberating about what play to run, then a value function that says "value winning over losing" will not provide any clarity unless the members of the team all agree about which set of actions are most conducive to winning. But this is exactly what I suggested was in dispute—the players disagree about which play (i.e. which set of behaviors) is most likely to lead the team to victory. Although there are different ways of specifying the collective value function, it cannot be underspecified; it must be complete over all of the outcomes under deliberation.

The preceding discussion captures something which is important in the rational choice with commitments model: commitments are important because they capture something like *intentions*. The football team is performing a collective action because they *share an intention* to run the play that they do, and sharing an intention is more important than sharing a judgment about reasons when it comes to collective action. The plural subject approach simply lacks the resources necessary to talk about intentions, and that lessens the explanatory power of that theory.

Another prima facie way that rational choice with commitments is explanatorily superior to plural subject theory is in the penumbral cases between individual and collective action. According to plural subject theory, either agents are engaged in team reasoning (in which case they are engaging in a distinct type of practical reasoning from individual reasoning) or they are not. Collective action requires that agents be engaged in collective reasoning. Thus, there should be no ambiguity about whether some action constitutes a collective action or a set of individual actions. For rational choice with commitments, however, the question is at least potentially less clear. According to rational choice with commitments, some behavioral profile b is part of c_b and not i_b so long as the agents performing b each share a commitment such that b is judged optimal according to their individual value functions. Commitments filter out certain outcomes from deliberation, and so the commitment might ensure that the agents coordinate on b by removing some more desirable options from some of their deliberation. (In PD games, the commitment to stay silent removes the defection option from deliberation, for example.) Agents who are maximally commitment rational will respect the commitment and play their part in b so long as doing so is most rational. However, the commitment need not *actually* filter any options out of deliberation at all. Suppose a set of agents shares a commitment to doing something that *they each individually judge they have most reason to do anyways*. This is something like Searle's example of the business school graduates who each believe that the world will most benefit if they each pursue their own selfish interests (Searle 1990). Assuming each graduate goes out and pursues her or his own selfish interest, is this a collective action?

Our intuition, I think, is that the answer to this question is not clear. A theory of collective action which reflects that unclarity is thus better than a theory of collective action which does not. For plural subject theory, there is no ambiguity allowed in answering that question. I think that Searle is correct to claim that if the business school graduates share some kind of *pact* to pursue their own selfish interests, that more clearly counts as a case of collective action. However, the emphasis on the 'pact' aspect of the action suggests that, in that case, the business school graduates recognize that they *might* be tempted at some point to not act in accordance with the agreement—that their commitment is, in effect, a modal commitment. In this case, the pact really is functioning as a commitment—it removes from rational consideration all those possibilities where agents are *not* selfishly pursuing their own interests. If, however, the commitment *always* lacks teeth, we might well wonder whether the agents are actually engaged in a collective action or a set of individual actions. If the collective goal never actually constrains anyone or affects their reasoning—if the goals of the collective *always* line up with the goals of all the individuals in the collective—then we might wonder whether there is any clear distinction to be drawn in this case between individual action and collective action. If I am right to suggest that our intuitions are not clear about what kind of action this is, then

that is a further reason to be suspicious of plural subject theory, since plural subject theory requires a clear answer to the question.

I do not take these arguments to be decisive against plural subject theory. I do think, however, that plural subject theory faces challenges. Commitments play an important role in facilitating collective action: they allow the members of a collective to express their own views about the best options for the collective, and indeed to try to persuade others in the collective to pursue some outcome rather than others, by ensuring that when it is time for the agents to deliberate and act, they are all willing to constrain their own judgments in deference to the judgment of the team. For the plural subject theorist, there is no diversity among the collective; everyone reasons exactly the same. This is an unrealistic view of collective action, but it is necessitated by the fact that a monolithic plural subject must be invoked in order to account for the difference between collective and individual action.

VIII. CONCLUSION

The aim of this paper was to show that RCT is incapable of distinguishing between two events which any adequate explanatory theory of action should be able to distinguish: namely, the difference between a collective action having been performed and an individual action having been performed. I rejected the view that the difference between the two lies in whether a cooperative outcome is brought about, since the same behavioral outcome might be brought about either through an individual or a collective action. Because we are interested in explaining actions, I argued that our explanation should be a maximizing explanation to at least some degree, and thus that we appeal to some kind of rational choice explanation. Given this, I argued that the best way to account for the difference within a rational choice framework is to find a way for our rational choice theory to be able to account for *commitments*, and that this is a more promising approach than the *plural subject* approach that has been advocated by some. If my arguments are sound, then the task still remains to develop such a model of rational choice with commitments and show exactly how such a model contributes to an explanation of collective action and individual action.[21]

ENDNOTES

[1] The term 'collective action' will not be precisely defined, but can be taken to refer to the distinctive type of action that agents engage in *together*: taking a walk together, playing music together, robbing a bank together. As Gilbert has pointed out, collective actions are often indicated by the appropriateness of the use of the pronoun 'we' by the participants (Gilbert 1996a).

[2] It is possible to accept Optimality and also allow for the possibility of intentional akratic actions; this usually involves a strict account of what it means to be in *possession* of a reason, such that an akratic agent can deliberate and act according to a reason that she is not in proper rational possession of. Under one plausible reading, this is the stance taken by Davidson (1969). Contrarily, we might just deny any link whatsoever between Optimality and akrasia, as for example by arguing that akrasia is a matter of the resoluteness of one's intentions (Holton 1999). It has thus been argued that one can accept Optimality and akrasia at the same time. Nevertheless, many skeptics of akrasia are motivated primarily by the plausibility of Optimality

and the implausibility of interpretations of the terms that allow Optimality to be compatible with akrasia.

[3] Compare Velleman 1992 and Frankfurt 1971.

[4] There is some question as to whether *satisficing* explanations are rational choice explanations, and if so, whether they are maximizing explanations. I will discuss this matter below, but my own view is that the two go hand in hand: if satisficing explanations are rational choice explanations, then they are also maximizing explanations (though the maximization may itself by constrained). If they are not maximizing explanations of any kind, however, then they are not rational choice explanations (though they might still speak to an interesting notion of rationality).

[5] Note that, at this level of generality, we do not need to specify what the value function is, or in virtue of what the agent chooses the optimal choice; nor do we need to specify what an agent can and cannot be. Thus, rational choice explanations can be given when the value function is unknown to the agent, or goes against the agent's own subjective value function (as, for example, when the value function is something like biological fitness); rational choice explanations can be given when agents' selection of the optimal outcome are not caused by conscious deliberation; rational choice explanations can be given when the 'agent' is an animal with limited deliberative capacities, or e.g. a hive of bees, or e.g. a firm, or e.g. a collection of neurons, etc. (see Ross 2005). As we will see, the fact that rational choice explanations can be given for firms or other collections of deliberative individuals does *not* yet resolve the question of whether rational choice explanations can distinguish between individual and collective actions.

[6] What I call behaviors are typically called 'actions' in the rational choice literature, and I will sometimes refer to them as such in this paper when it does not lead to confusion. However, we should be careful to distinguish between the two. The fact that some behavior is optimal according to RCT does not mean that the performance of that behavior is necessarily an action, and when we want to provide an explanation of what makes some behavior an action, we must appeal to something other than RCT. To take a trivial example, an agent can perform an optimal behavior while sleepwalking, or through a physical spasm; the fact that the behavior is optimal does not suffice to make it an action. In order for RCT to contribute to an intentional explanation of action, the agent must be aware of the optimality and choose the behavior *because* it is optimal.

[7] Because we are taking decision problems to be relative to individual agents—G is a decision problem *for agent* X—a number of counter-intuitive results follow, especially for those accustomed to modeling objective decision problems. For example, because agents might hold false beliefs about the actions available to other agents, the agents might jointly bring about an outcome which is not represented in *any* of their subjective decision problems (and which therefore is not represented in the value ordering). Though counter-intuitive, these results do not, I think, pose any serious threat to the desirability of using such subjective decision-theoretic models for the current project.

[9] We also use the term 'parametric choice' to describe scenarios in which there is more than one agent, but where every agent except one chooses an action independent of the choices available to the other agents. In other words, there are many people doing things, but only one chooses what to do on the basis of what others *might* do. This is subsumed under my usage; I

hold that there is only one agent choosing among a set of behaviors, and the outcome produced is a function of only one agent's choice.

[10] A caveat: behavior compositionality does *not* entail the claim that the collective *action* is fully determined by the constituent individual *actions*. Behavior compositionality is a claim about behaviors (i.e. bodily movements) only; it makes no claims about the actions those behaviors help constitute.

[11] By 'justify' here I mean nothing more than that the agent has subjective reasons as described at the outset of the paper.

[12] The argument that I present here is general, and applies to *all* sets of behaviors. We might worry that this is implausibly strong, and that there are some sets of behaviors which either cannot possibly support a series of individual actions, e.g. two people dancing the tango together, or cannot possibly support a collective action, e.g. one person stabbing another person in the back. I think we have good reason to believe that the argument ought to be general, but nothing hinges on this; we can restrict the argument to just that set of behaviors which admits of possible ambiguity between individual and collective actions.

[13] Although traditional RCT typically defines preferences over behaviors (and hence rational choice as only applying to the selection of strategies), I will speak interchangeably here about the rationality of *performing behaviors* and the rationality of *bringing about outcomes*. In order to do this, of course, we need to specify a probability function which maps behavioral profiles onto outcomes. It is worth emphasizing, however, that preferences over behaviors presuppose a consistent underlying ranking of outcomes, and that when behavior b_1 is preferred over behavior b_2, this is because the expected utility of b_1 is higher than that of b_2—and this is because the outcomes that b_1 and b_2 produce can themselves be compared for desirability.

[14] The clearest presentation of this view of intentions as rational constraints on deliberation and future action can be found in Bratman 1987.

[15] As Bratman (1984) has pointed out, the intention need not be an *intention to* b.

[16] An objection could be made that the preceding argument is too broad. After all, I want to argue that we need a theory of rational choice with commitments in order to account for collective action. Giving an explanation of both individual and collective actions requires an appeal to intentions, however, and so one might think that we need a theory of rational choice with commitments in order to account for individual action as well as collective action. In fact, I will argue in section VI that there is an asymmetry between the two cases—the individual action cases are different in important respects from the collective action cases. Because of these differences, I will argue, we do not need a theory of rational choice with commitments in order to explain individual action, though we do for collective action.

[18] The latter clause here covers cases in which only one agent chooses rationally, and other agents are 'choosing' without any regard to the probable choices of the other agents.

[19] This is not to say that X's value function is incapable of being other-regarding, of course. X's value function might e.g. rank outcomes in which her friends' preferences are satisfied more highly than outcomes in which her friends' preferences are not satisfied. See Sen 1977.

[20] The term 'plural subject theory' comes from Gilbert; in the game theoretic literature, the phrase 'team reasoning' is more often used. There are notable differences between the two approaches, including the issue of the ontological status of the entities being studied. In spite of

the differences, it is useful to consider the two approaches together here, as there are significant similarities between the two.

[21] Thanks to Terence Cuneo, Tyler Doggett, Kareem Khalifa, and the members of the Ethics Reading Group at the University of Vermont for helpful comments on an earlier draft of this paper.

BIBLIOGRAPHY

Alonso, Facundo. 2009. “Shared intention, reliance, and interpersonal obligations.” *Ethics* 119(3): 444–475.

Bacharach, Michael. 1999. “Interactive team reasoning: A contribution to the theory of co-operation.” *Research in Economics* 53(2): 117–147.

———. 2006. *Beyond individual choice: teams and frames in game theory*. Princeton, N.J: Princeton University Press.

Bardsley, Nicholas. 2007. “On collective intentions: collective action in economics and philosophy.” *Synthese* 157(2): 141–159.

Bratman, Michael E., et al. 1988. “Plans and resource-bounded practical reasoning.” *Computational intelligence* 4(4): 349–355.

Bratman, Michael E. 1984. “Two faces of intention.” *The Philosophical Review* 93(3): 375–405.

———. 1987. *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

———. 1993. “Shared intention.” *Ethics* 104(1): 97–113.

Cohen, Philip R. and Hector J. Levesque. 1990. “Intention is choice with commitment.” *Artificial Intelligence* 42(2-3): 213–261.

Elster, Jon. 2000. *Ulysses unbound: studies in rationality, precommitment, and constraints*. Cambridge: Cambridge University Press.

Frankfurt, Harry G. 1971. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68(1): 5–20.

Gauthier, David. 1986. *Morals by agreement*. Oxford: Clarendon Press.

———. 1996. “Commitment and choice: an essay on the rationality of plans." In *Ethics, Rationality, and Economic Behaviour*, ed. Francesco Farina, et al. Oxford: Oxford University Press, 217–243.

Gilbert, Margaret. 1989. *On social facts*. London: Routledge.

———. 1996. *Living together: rationality, sociality, and obligation*. Lanham: Rowman & Littlefield Publishers.

———. 2001. “Collective preferences, obligations, and rational choice.” *Economics and Philosophy* 17(1): 109–119.

———. 2003. “The structure of the social atom: Joint commitment as the foundation of human social behavior.” In *Socializing metaphysics: the nature of social reality*, ed. Frederick Schmitt. Lanham: Rowman & Littlefield, 39–64.

———. 2006. “Rationality in collective action.” *Philosophy of the Social Sciences* 36(1): 3– 17.

Gold, Natalie and Robert Sugden. 2007. “Collective intentions and team agency.” *The Journal of Philosophy* 104(3):109–137.

Holton, Richard. 1999. “Intention and weakness of will.” *The Journal of Philosophy* 96(5): 241–262.

Meacham, Christopher J. G. and Jonathan Weisberg. 2011. “Representation theorems and the foundations of decision theory.” *Australasian Journal of Philosophy* 89(4): 641–663.

Rao, Anand S. and Michael P. Georgeff. 1991. “Modeling rational agents within a BDI-architecture.” Technical report, Australian Artificial Intelligence Institute.

Ross, Don. 2005. *Economic theory and cognitive science: microexplanation*. Cambridge, MA: MIT Press.

Roth, Abraham S. 2004. “Shared agency and contralateral commitments.” *The Philosophical Review* 113(3): 359–410.
Roy, Olivier. 2010. “Interpersonal coordination and epistemic support for intentions with we-content.” *Economics and Philosophy* 26(03): 345–367.
Savage, Leonard J. 1954. *The foundations of statistics*. New York: Wiley.
Searle, John R. 1990. “Collective intentions and actions.” In *Intentions in communication*, ed. Phillip R. Cohen et al. Cambridge, MA: MIT Press, 401–415.
Sen, Amartya. 1977. “Rational fools: A critique of the behavioral foundations of economic theory.” *Philosophy and Public Affairs* 6(4): 317–344.
———. 1985. “Goals, commitment, and identity.” *Journal of Law, Economics, & Organization* 1(2): 341–355.
———. 2005. “Why exactly is commitment important for rationality?” *Economics and Philosophy* 21: 5–14.
Simon, Herbert A. 1959. “Theories of decision-making in economics and behavioral science.” *The American Economic Review* 49(3):253–283.
Sugden, Robert. 1991. “Rational choice: A survey of contributions from economics and philosophy.” *The Economic Journal* 101(407): 751–785.
Sugden, Robert. 2000. “Team preferences.” *Economics and Philosophy* 16(2): 175–204.
Tuomela, Raimo. 2005. “We-intentions revisited.” *Philosophical Studies* 125(3): 327–369.
van Hees, Martin and Olivier Roy. 2008. “Intentions and plans in decision and game theory.” In *Reasons and Intentions*, ed. Bruno Verbeek. Aldershot: Ashgate, 207–226.
Velleman, J. David. 1992. “What happens when someone acts?” *Mind* 101(403): 461–481.
———. 1997. “How to share an intention.” *Philosophy and Phenomenological Research* 57(1): 29–50.
Von Neumann, John and Oskar Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Wooldridge, Michael J. 2000. *Reasoning about rational agents*. Cambridge, MA: the MIT Press.