Commitment as Motivation:
Amartya Sen's Theory of Agency and the Explanation of Behavior[1]

Ann E. Cudd, University of Kansas

Abstract:

This paper presents Sen's theory of agency, focusing on the role of commitment in this theory as both problematic and potentially illuminating. His account of some commitments as goal-displacing gives rise to a dilemma given the standard philosophical theory of agency. *Either* commitment-motivated actions are externally motivated, in which case they are not expressions of agency, *or* such actions are internally motivated, in which case the commitment is not goal-displacing. I resolve this dilemma and accommodate his view of commitment as motivation by developing a broader descriptive theory of agency, which recognizes both agent goal-directed and goal-displacing commitments. I propose a type of goal-displacing commitment, which I call "tacit commitment," that can be seen to fit between the horns. Tacit commitments regulate behavior without being made conscious and explicit. This resolution suggests a means of bridging the normative/descriptive gap in social-scientific explanation.

1. Introduction

Amartya Sen is the most important and prolific living philosopher-economist, having made seminal contributions to economic science in the fields of social choice

[1]An earlier version of this paper was presented at a Symposium on the Work of Amartya Sen, Erasmus University, Rotterdam, The Netherlands, July 1, 2010, and I am especially grateful to Ingrid Robeyns, who organized the symposium, and Amartya Sen, who commented on the papers, as well as my co-symposiasts Mozaffar Qizilbash, Henry Richardson, and Ingrid Robeyns, and the audience for the session. I also thank Elizabeth Anderson and Neal Becker for feedback on the earlier version. The paper benefitted from discussion with audiences at the University of Washington and the University of Kansas Departments of Philosophy. Finally, I received excellent suggestions from two anonymous referees and from the Editor of *Economics and Philosophy*, for which I am very grateful.

theory, welfare economics, feminist economics, and the explanation of famines,[2] but also original contributions to consequentialist ethics, political philosophy, feminist philosophy, identity theory, and the theory of justice. He draws many deep connections between the descriptive and normative approaches in both economics and philosophy. In Sen's view, economic science must comprise normative as well as descriptive elements in order to successfully explain human behavior. Human agents respond to many kinds of incentives and constraints that are normatively derived or inflected. However, normative and descriptive approaches sometimes conflict and give rise to contradictory ideas, as can be seen in Sen's work on agency and commitment.

Although he has written extensively on agency in the context of moral and political theory,[3] he has been less explicit about his descriptive theory of agency for use in explanatory theory. This can in part be explained by the fact that neo-classical economists tend either to downplay or eliminate the role of agency, or to give it a simplistic motivational structure in their explanations of behavior. Sen rejects both the behaviorist and reductionist impulses, however. His writings offer an illuminating descriptive theory of agency, which attributes two very different types of motivations to behavior: self-interest and other-directed commitment. Sen, as I shall argue, ultimately grounds agency not in intentional goal-directed behavior, but in a broader view of norm-governed behavior.

In this paper I will present Sen's theory of agency, focusing on the role of commitment in this theory as both problematic and potentially illuminating. Sen's discussion of commitment begins as part of his rejection of the behavioristic foundations of rational choice theory begun in his early work. This critique was initially seen as being aimed only at revealed preference theory, which made some commentators think that the critique could be rather easily incorporated into economic theory in a way that affected it minimally. However, the critique goes deeper than that. In fact, Sen recognized a whole category of human motivation – which he calls "commitment" – that neither revealed

[2] Sen is the 1998 winner of the Nobel Prize in Economic Sciences.

[3] See for examples (Sen 1982, 1985b, 1999).

preference theory nor a less behavioristic interpretation of standard preference theory can easily accommodate, if at all.

Commitments, according to Sen, may replace or even displace the agent's own goals. Sen's account of commitment as goal-displacing gives rise to a dilemma, however, given the standard philosophical theory of agency. *Either* commitment-motivated actions are externally motivated, in which case they are not expressions of agency, *or* such actions are internally motivated, in which case the commitment is not goal-displacing. I resolve this dilemma and accommodate his view of commitment as motivation by developing a broader descriptive theory of agency, which recognizes both agent goal-directed and goal-displacing commitments. This resolution suggests a means of bridging the normative/descriptive gap in social-scientific explanation.

2. Sen's normative account of agency and its implications for the descriptive account

For Sen, an agent is ''someone who acts and brings about change, and whose achievements can be judged in terms of her own values and objectives, whether or not we assess them in terms of some external criteria as well.'' (Sen 1999:18). Agency involves two discrete elements: formulating an end and acting on or pursuing that end. One of the unique contributions of Sen's normative account of agency is to recognize that persons have two aspects that need to be considered in moral and political theory: an agency aspect, which is "the moral power to have a conception of the good" (Sen 1985: 186) and a well-being aspect, which consists, roughly speaking, of the things that make a life objectively go well for a person.[4] Any moral theory that is concerned about the consequences of actions or policies on people's lives must consider persons' well-being, but such theories must also be concerned about the persons' own intentions and desires, independently of how the actions that they give rise to affect persons' well-being. To be an agent is to form a conception of the good, which may involve raising one's level of well-being, but may also at times involve sacrificing one's well-being for something else that one values. "Some types of agency roles, e.g., those related to fulfilling obligations,

[4] On Sen's view well-being is to be understood in terms of capabilities for functioning see (Sen 1999).

can quite possibly have a negative impact on the person's well-being. Even when the impact is positive, the importance of the agency aspect has to be distinguished from the importance of the impact of agency on well-being." (Sen 1985: 187)

This normative account of agency determines some elements of a corresponding descriptive account, but leaves others indeterminate. Descriptively, it implies that agents formulate their values and objectives and act in light of those values and objectives, which may not be connected with the person's own well-being in any sense. Nonetheless, the motivations for action – the values and objectives – belong to the agent. Thus, agents define their own ends. The distinction between the agency aspect of persons and the well-being aspect of persons is one that does not fit well in the standard descriptive model of agency. Well-being, on this model, is part of the utility function of the agent insofar as she ranks it in her preference ordering. That model subsumes well-being entirely within the agency aspect, and leaves any aspect of agents' objective well-being that does not play a role in the agent's own preference ordering with no explanatory role. Thus one task of the descriptive theory of agency is to discover and explain agents' objectives. Furthermore, the normative theory of agency does not determine how values and objectives give rise to actions. The economic theory of behavior explains behavior as the rational pursuit of agents' values and objectives.

3. Complicating the descriptive model of agency: context dependence and commitment

The standard neoclassical model of economic agency holds that agents are consistent preference bearers who make choices among their available options to maximize the satisfaction of their preferences. Preferences are represented by utility functions, and choice is represented by mathematical maximization of those functions subject to constraints. Some insights into Sen's descriptive theory of agency can be gleaned from his critique of revealed preference theory. Sen's critique of preference theory can be characterized generally as having two broad targets. The first, his critique of revealed preference theory, aims to refute the behaviorist attempt to derive preference from choice behavior. The second, his theory of commitment as an alternative source of motivation, aims to undermine the notion that rational behavior must be self-interested and to provide an alternative formalization of rational motivation. The two aims are

related, and he treats both of them in his influential (1977) “Rational Fools” paper, although some other papers address separately one or the other.

3.1. Rejecting revealed preference theory

In order to even get to the point of discussing agency, intention, and motivation as part of economic theory, Sen first had to critique the behaviorism of revealed preference theory, which purports to avoid all that by deriving preference from observations of individuals’ externally verifiable choice behavior. Influenced by the behaviorist commitment to limiting models to observables, economists beginning with Samuelson in the mid-20th century developed revealed preference theory (RPT), which derives preference from choice behavior, which is, at least in principle, observed. Revealed preference theory is based on the simple and operational idea that if an agent chooses x when y is available (within his budget constraint) then he prefers x to y. Arrow formalized this as the Weak axiom of revealed preference (WARP): if an agent prefers x to y, then she must not choose y over x when both are available to her. WARP thus places an internal consistency requirement on choice; rational agents are assumed to adhere to this requirement, and hence preference can be inferred from observed choices. In his (1973) “Behavior and the Concept of Preference” paper, Sen points out that this theory allowed Arrow to derive consumer behavior theory “with economy” and opened up empirical studies of preference by observing market behavior.

Despite its “economy,” Sen criticized this theory on several grounds. First, and most fundamentally, it mystifies the causal order of preference and choice. Arrow aimed to free theory from mentalistic conceptions of preference and utility, because of longstanding concerns about the unobservability of mental states. But Sen argues that the theory is useful only under the assumption that there are preferences in the head of the agent that adhere to the reasonable assumption of WARP. Realistically speaking, from the point of view of the agent the preference comes first, although from the point of view of the scientist, preference is inferred from choice. Sen thus distanced himself from the behaviorist impulse to give a reductive account of agency, allowing for a more complicated and realistic theory of choice behavior and motivation.

Second, Sen argued that choices can appear to violate the internal inconsistency requirements because of their context dependence, which can include menu dependence,

strategic opportunities, the situational and cultural dependence of social norms, and identities of the agents involved. In different articles over the course of several decades, Sen presents examples of how the choice act injects additional elements that make the internal consistency of choice impossible to observe directly from the choice itself. One is what he calls the chooser dependence of the choice, and this is illustrated by the mango example: a group of people each pick from a basket of apples and mangoes; having the last pick of fruit from a basket that includes one mango is a different choice situation than having the second to last pick when there is only one mango left. Even if one prefers the mango all other things equal, one might not wish to be seen as greedy or selfish by choosing the last mango and depriving another of that opportunity. Observing simply the choice behavior of an agent in the two situations leads one to the conclusion that the agent both prefers mangoes to apples and apples to mangoes. By recognizing how social norms contextualize the choice, the apparent inconsistency of the preferences can be explained away. And this means that social norms provide agents with another source of motivation that the theory needs to be able to capture.

Sen argues that because the choice act itself is meaningful to agents, revealed preference theory cannot serve the behaviorist goal of avoiding getting into the heads of agents, or the positivistic role of avoiding positing realistic assumptions about how agents think about choices. In this way, social science models, while descriptive, are essentially different from those in the natural sciences because the former must take facts about norms into account. Sen points out a difference between maximization problems in physics and in descriptive sciences of human behavior. In the latter there is the fact of volition "maximizing behavior differs from non-volitional maximization because of the fundamental relevance of the choice act." The choice act is important because it situates the chooser in a social context of norms that make the choices meaningful to agents and constrain the ways in which they choose beyond the physical constraints of the material situation at hand.

Partly due to Sen's work, the standard economic model of agency has been made more complex and enhanced by the recognition of informational asymmetries and the possibility of making strategic choices in contexts of interaction among other rational agents. Sen's argument against revealed preference theory involves showing different

types of interests that a person might be satisfying with a given choice, including immediate self-interest, strategic interests, or an interest in adhering to moral duties or social norms. Hence, a single-valued utility function that is equated with any one of these interests will not be explanatorily sufficient, and *a fortiori* there is no simple maximization of such a function that will be sufficient.

3.2. Self-interest vs. commitment

Sen's critique of the behavioral foundations of economic theory sets off from the critique of preference as inferred from choices and internal consistency conditions, but it does not stop there. The next step is to critique the assumption that all rational action is self-interested. One of the most important contributions to clarifying the economic theory of agency has been to analyze the assumption of self-interested behavior into three separable aspects, as follows:[5]

1. *self-centered welfare* – the assumption that a rational person's welfare depends only on her consumption.

This is the assumption that agents value only their own consumption. On a restrictive notion of well-being as dependent only on the agent's consumption of goods, self-centered welfare is the assumption that the agent's self-interest is equivalent to her well-being. This assumption states how agents formulate their values, not their goals for acting or how they act in light of those goals.

2. *self-welfare goal* – the assumption that a rational person's goal is to maximize the expected value of her own welfare.

This assumption states that an agent's goal in acting is to maximize only her own objectives. It is not about what constitutes welfare or what the agent values, which could be narrow well-being, as with assumption 1; but welfare could also include another's well-being. Nor is it about how agents choose in light of their goals, which is covered by the third aspect of self-interest.

3. *self-goal choice* – the assumption that the rational person maximizes the satisfaction of her goals irrespective of others' goals.

---

[5] (Sen1987) is I believe the first time he introduced this three-way distinction.

This assumption is about how agents' choose, namely, to achieve their own goals. The goal might include raising another's welfare, or indeed, any other sort of goal, including collective goals, moral goals, or group norms *as long as the agent takes that on as her goal*.

Although all three aspects are assumed in some economic models, self-centered welfare is fairly easily and readily given up. It is not only extremely unrealistic to assume that people do not get any welfare from anyone else's consumption bundles, it is also unnecessary for the formalization of most of economic theory.[6] There may be some good reasons to incorporate this assumption into one's ethical theory,[7] but descriptively it seems clearly false and misleadingly so.[8] It would not allow us to make sense of any sort of other-directed behavior, whether positive or negative. The assumption of self-welfare goal is more commonly made and defended, but Sen questions this assumption for explaining some actions.

In his "Rational Fools" paper Sen introduces the distinction between sympathy and commitment, two types of other-directed motivations. Sympathy involves one's own feelings about the experiences of others; it is "the case in which concern for others directly affects one's own welfare." (Sen 1977: 95) Thus, admitting sympathy as a type of motivation, while it clearly does violate self-centered welfare, does not violate the self-welfare goal assumption. Commitment as a motivation, however, drives a wedge between

---

[6] If self-centered welfare is given up along with self-welfare goal the resulting theory of motivation cannot be so easily incorporated into formal economic theory.

[7] The assumption of self-centered welfare implies that agents do not take an interest in the interests of others, an assumption called "non-tuism." This assumption is useful in answering the moral skeptic because it sets the bar very high for showing why someone still ought to be moral. See for example (Dimock 1999).

[8] Keith Dowding (2002) seems to disagree with the "misleading" part of this claim, though, because he thinks that when used in aggregate models, the self-centered welfare goal is approximately correct and will lead to correct estimates of aggregate demand functions. For the moment I confine the discussion to the understanding and assessment of individual preferences.

personal welfare and choice. When one acts on a commitment, one does something for the sake of a principle, a promise, a group norm, or the anticipation of future welfare, but not – at least not directly – for one's own welfare as one presently conceives it. Sen offers examples of choices where commitment seems to be at work including: contributing to public goods; truthfulness (i.e., that people normally tell the truth rather than simply say whatever is in their self-interest to say); voting; fiduciary responsibility; environmentally-sensitive conduct, and work motivation.

Although "commitment" has a positive connotation in ordinary speech, Sen does not assume that all behavior that takes commitment as its goal is positive let alone in the interest of all. "Groups intermediate between oneself and all – such as class, community, or occupation groups – provide the focus of many actions involving committed behavior." (Sen 1987: 20) Group loyalty may involve sacrifice of one's "purely personal" interests for the sake of the group or the group's cause, but the cause may or may not be a morally good one.[9] "The mixture of selfish and selfless behavior is one of the important characteristics of group loyalty, and this mixture can be seen in a wide variety of group associations varying from kinship relations and communities to trade unions and economic pressure groups." (Sen 1987: 20) Some forms of commitment are morally suspect because of the way that the self is sublimated, as often happens with oppressed persons who come to believe that they are naturally or deservedly treated as second-class citizens.[10]

It might be objected here that such explanations of behavior can be reduced to explanations in self-centered welfare terms. For example, acting on a promise might be explained as acting on one's desire to appear to be a promise-keeper because, by appearing trustworthy and therefore being included in interactions that increase one's

---

[9] Andrew Oldenquist (1982: 176) discusses loyalty as a third form of normative reason apart from self-interest on the one hand and impartial morality on the other. Loyalties on his account are similar to some forms of commitments on Sen, although I am discussing commitments here as descriptive motivations, not necessarily justifications of behavior.

[10] I thank Neal Becker for pressing this point, which I have written extensively about in (Cudd 2006).

future consumption, that raises one's own (self-centered) welfare. But this reductivist impulse conflicts with our subjective feeling that in some cases it is not a desire but rather some external reason that motivates us. Furthermore, once the behaviorist impulse to avoid reference to internal mental states has been abandoned, there is no theoretical reason to impose either self-centered welfare or self-welfare goal assumptions. As long as commitments are not seen as violating self-goal choice, the theory of rational action as maximization of some objective function, some goal of the agent, still applies. The question is what goal is being maximized by the agent and whether this goal is internally or externally generated? Thus, there is no theoretical reason to reduce commitment to some basic desire, and retaining commitment as an explanatory variable better accommodates our intuitions and subjects' reports in a wider variety of situations.

3.3. Commitment, rationality, and agency

Just how radical is the addition of commitment to the descriptive theory of agency? That depends upon whether the commitment can be seen as adhering to the assumption of self-goal choice. According to recent interpretations of Sen's view, acting from the motivation of commitment may or may not involve violation of self-goal choice. In order to explicate Sen's view, Philip Pettit (2005) helpfully distinguishes between goal-modifying commitment and goal-displacing commitment. A goal-modifying commitment is a commitment that alters the agent's own goals based on recognition of others' goals and of how the agent's behavior affects them. A goal-displacing commitment is a motivation that replaces – as the determinant of the agent's choice – the agent's goals with another's goals or the goal of a group, or possibly an impartial moral norm. It is only this latter sense of commitment as goal-displacing that violates the assumption of self-goal choice.

The assumption of self-goal choice is deeply connected with the neoclassical economists' understanding of rationality, but also, as Pettit clearly points out, with the dominant philosophical understanding of agency. On this understanding, an agent just is a being that acts to achieve *its* goals, where the possessive is as important as the existence of a goal toward which the act is directed. The agent's goals might of course involve the goals of a group that the agent identifies with; as Sen states, "'We' may be the natural unit of first-person decision." (Sen 1986b: 351) But acting from a collective goal that one

takes as at least in part one's own is not mysterious. What would be puzzling is *action* which is not motivated by a goal with which the agent somehow identified or takes as her own. One who does not act towards one's own goals is not autonomous in the most minimal sense of the term; there is something robotic, slavish, or simply non-agential about such a being. Now there are several things to be said about this point, and I will get to them in the next section when I explore varieties of committed action. But first I want to explore further the idea that commitment violates the economists' understanding of rationality.

In his early work on the behavioral foundations of economic theory in the 1970s, Sen takes maximization of goals to be the primitive notion of rationality, while relaxing the notion that these goals had to be self-interested. Sen argues that the real problem is the idea that an agent's motivations can be expressed by one, univocal function. "The *purely* economic man is indeed close to being a social moron. Economic theory has been much preoccupied with this rational fool decked in the glory of his *one* all-purpose preference ordering." (Sen 1977: 102) Sen proposed ways of describing non-self-interested choices as rational in the sense of maximizing by formalizing a model that allows a dual or multivalued preference ordering. In his (1970) paper "Choice, Orderings and Morality," Sen is concerned to show how mutual non-confession could be seen as rational in the Prisoner's Dilemma (PD). Here he proposes that we view the agents as facing a choice between rank orderings of the strategy pairs based on PD type preferences (where each has a dominant strategy to defect) and Assurance Game (AG) preferences (where the first choice of each is to cooperate if the other would as well, although they prefer defection if the other might defect) and Other Regarding (OR) preferences (where each unconditionally prefers to cooperate with the other). If agents act as if they have AG or OR preferences, then they will do better than if they act as if they have PD preferences, even if they face a PD in terms of their individual welfare in the situation.

Sen generalizes this to discuss the orderings of orderings, and in "Rational Fools," he further refines the idea of meta-rankings of preference orderings. A meta-ranking is a ranking according to the preferences one has for particular preference orderings. Let X be the set of alternative and mutually exclusive combinations of actions under consideration, and let Y be the set of rankings of the elements of X. A ranking of the set Y will be called

a meta-ranking of action set X. Choosing according to meta-rankings will allow the model to express commitments to a variety of things, such as deontic obligations, morality, or other non-self-interested principles, by the meta-ranking, which can then be treated as a preference ordering that can fit into a maximizing model. Sen points out that this account of meta-rankings is a "structure", not a "theory", and the structure allows multiple theories. In sum, he writes, "the apparatus of ranking of rankings assists the reasoning which involves considering the merits of having different types of preferences (or of acting as if one has them)." (Sen 1977: 107)

This work was prior to his clearly drawing the three-way distinction that separated self-goal choice from the other two aspects of self-interest, and prior to Pettit's useful distinction between types or conceptions of commitment. It is these latter distinctions that permit us to clarify just how radical Sen's account of commitment really is by asking whether this account includes goal-displacing commitments as well as goal-modifying ones. Armed with this distinction, we can read the phrase in the previous quote "having different types of preferences" as essentially different from the phrase "acting as if one has them," in that the former is consistent with goal modifications, but the latter is not (if one does not in fact have those preferences). If goal-displacing commitment is to be understood as a way of "acting as if" one had preferences other than one in fact has, then this meta-ranking preference model can provide a formalization of goal-displacing commitment as action that is explained as maximizing some objective function.

It may seem doubtful to suggest that Sen would insist that the model of rational behavior as maximization be retained when commitment is appealed to as the explanatory motivation. Writing on Sen's conception of preference, Elizabeth Anderson (2001) interprets Sen's theory of commitment as an account of motivation (i.e., motivation by a moral principle or a social norm of responsibility) that cannot be captured by a principle of maximization of utility even when utility is taken in the widest sense of whatever one values. She summarizes her exposition of Sen: "Thus, for explanatory purposes, Sen instrumentalized the concept of preference in two ways: first, by disambiguating the concept, replacing it with three distinct concepts (choice, underlying motive, and welfare), and second, by articulating an alternative model of behavior, commitment, that was not framed in terms of preference satisfaction at all." (Anderson 2001: 23) However,

Anderson sees this as implying a lacuna in his work, since he does not propose an alternative to the maximizing framework that would enable us to conceive of committed action as rational. "Sen does not propose an alternative, non-preference-based conception of rationality in terms of which committed action makes sense." (Anderson 2001: 24) Sen's response was to say that she misunderstood his claims about the theoretical reach of the maximizing framework, suggesting that the maximizing framework indeed could be stretched to rationalize goal-displacing commitments. (Sen 2001: 57) But even if the maximizing framework can be extended to account for goal-displacing commitments, Anderson still has a point about the conception of rationality in terms of which goal-displacing committed action makes sense. In particular, we are owed an explanation of *why it is rational to act as if one has goals that one does not in fact take as one's own*.

Perhaps it is too quick to infer that Sen intends to subsume goal-displacing commitments under the maximizing framework, however. In a 1986 article, Sen offers a different account of rationality of choice, which characterizes rationality as whatever is not in the rejection set R(S) of those options that the agent decides are, on reflection, not to be chosen. He calls this notion of rationality "correspondence-rationality," and says about it specifically "reflective choice is not required to correspond to the maximization of some particular thing." (Sen 1986b: 346) Indeed he says that "there might not even be any 'everything considered' maximand."(347) So this leads me back to reading Sen as Anderson, Pettit, and others have read him, as making the more radical claim that goal-displacing commitments cannot be explained within the maximizing framework. (Anderson 2001; Pettit 2005; Hausman 2005) Instead, they may be subsumed under this notion of correspondence-rationality, which covers various kinds of reflective thought that can be considered rational, but that are not further specified by Sen. The fundamental issue, though, is not whether goal-displacing commitments fit into the maximizing framework of rational choice theory. Rather it is an issue about agency. Is it an expression of agency to act on a commitment that one does not take as one's own?

4. Using commitment to explain behavior

The essential thing for interpreting Sen's descriptive theory of agency is to see how commitments play a role in the explanation of behavior. Five different types of

explanatory models that make use of commitment can be distinguished. The first three of these can be accommodated within the standard theory that takes agents to be maximizers of some objective function, but the latter two involve goal-displacing commitments that cannot fit this theory.

First, commitment can be seen as a means of building a reputation or as a way of restraining oneself. In explaining behavior this way, commitment adheres to the assumptions of self-centered welfare (by helping to build a reputation to promote long-term consumption) and self-welfare goal. Even if the motivation is derived from a social group norm, it individualizes the group motivation by modeling the agent as using the commitment instrumentally to further her own interest.[11] Economic theories recognize this as a form of signaling behavior when it occurs in strategic contexts, and there is nothing radical in the use of this type of explanation in economics. (Binmore 2005) Commitment can also be used as a pre-commitment of the self in non-strategic contexts, e.g., to control one's addiction to cigarettes by throwing away the full pack.[12]

Second, persons sometimes choose to act on a principle or norm that they have deliberated on but that does not necessarily involve their own consumption. This intentionally committed behavior adheres to the assumptions of self-welfare goal as well as self-goal choice, but not self-centered welfare. Modeling behavior this way internalizes the external motivation of commitment, and is typical of identity formation. (Sen 1985a, 2006) The standard economic model still individualizes the group motivation by taking the group goal as the individual's own, and it expresses the motivation as rational within the maximizing framework. Models of Rousseau's stag hunt or Sen's Assurance Game can be seen as expressing this, because the individual orders her preferences to rank the cooperative strategy above defecting, but the individual's own ranking still determines the rational action. Such commitments may not replace all the

[11] This is not the form of commitment that Sen had in mind when he introduced the term, but such a reductivist account is possible. See (Gauthier 1986; Morris 2010).

[12] Elizabeth Anderson pointed out to me this non-strategic use of commitment. See also (Gauthier 1996).

other goals that an individual has, of course, and an individual may strategically pretend to be seeking this group goal when in fact she is shading toward her own.

Third, persons can act out of social commitment and moral imperatives that they embrace as at least in part their own goals. Sometimes agents choose based on principles that are recognized in their communities, or that they recognize, as socially beneficial or morally good. If these commitments were seen as shaping the ultimate goal the agent seeks, then they could be subsumed under the maximizing framework by modeling the objective function to be maximized as reflecting these all-in goals. (Bossert and Suzumura 2009) This model of behavior involves denial of self-welfare goal but still adheres to self-goal choice.

Fourth, persons can act, according to Sen, on commitments that replace their own goals. Sen's primary example of such reasoning invokes the Prisoner's Dilemma. The dilemma for the game theorist is how to explain the fact that people often cooperate in PD situations though standard game theoretic rationality requires defection. The various game-theoretic explanations and evasions of the problem reject the possibility of explaining cooperation in a true one-shot PD as rational. Here is where goal-displacing commitment plays a crucial role for Sen. By recognizing that the only way to "solve" the PD is to commit to cooperate, despite the fact that doing so entails sublimating their own goals to that of the community, persons can act rationally on a commitment to a goal that is not their own. Why do people do this, according to Sen? It is a kind of social thinking, "part of living in a community." (Sen 1985a: 212) Modeling persons as behaving in this way preserves the notion that players have mutual knowledge of the game situation, and that the game accurately represents the individual preferences as the typical PD-type preferences.[13] But it represents the players as playing as if they are playing an assurance game. Behaving 'as if' they have assurance game preferences can be interpreted as a way of adopting a group identity. Thus, according to Sen, persons can act rationally on group-based preferences that are not their own but from which they rationally act as if they

[13] In preserving mutual knowledge of the game situation, it is unlike the Kreps, et.al. (1982) type of solution of the PD. In preserving the individual preferences it is unlike solutions that suppose that the individual mistakes the one-shot PD for a repeated game.

were. We can answer Anderson's question in this case by pointing out that it is rational to act on these preferences that are not one's own because by doing so one does better according to one's own preference ordering than if one acts on those preferences directly.[14]

Finally, behavior can be explained based on conventional rule following. Insofar as this is to be distinguished from explaining behavior as acting from social commitment and moral imperative, this explanatory schema also allows for behavior as motivated by a rule that displaces the agent's own goal. Such behavior can be explained by evolutionary game theory, which is different from an intentional explanation with behavioral rules being deliberately chosen by an individual who considers how they should act. Sen points out that the evolutionary explanation can be combined with any of the other explanations because long-run survival might be enhanced by the ability to consider behavior in this way, that is evolutionary processes might affect not only the rules of behavior but also our psychological preferences. Or, I might add, our social norms or conventional rules.

It is these last two types of explanatory models invoking commitment that call into question our usual understanding of agency as action based on the agent's own goals. Sen claims that there is an "essential and irreducible" duality in the conception of persons as agents with goals and commitments, who also have a well-being that calls for attention.

> This dichotomy is lost in a model of exclusively self-interested motivation, in which a person's agency must be entirely geared to his own well-being. But once that straitjacket of self-interested motivation is removed, it becomes possible to give recognition to the indisputable fact that the person's agency can well be geared to considerations not covered

[14] As an anonymous referee pointed out, this response to Anderson raises a paradox related to the paradox of hedonism, that one best achieves happiness by not pursuing it directly. My main concern is the problem of asserting of an agent that she is pursuing a goal that she does not actually have, but merely acts as if she has. I will argue in the next section that acting on a tacit commitment or internalized social norm is a type of agency that allows the agent to be seen as pursuing and yet not aiming at a goal.

> – or at least not fully covered – by his or her own well-being. (Sen 1987: 49)

Sen calls these the "well-being aspect" and the "agency aspect" of persons, as was noted earlier. Failure to recognize this duality of persons, Sen claims, impoverishes both the normative and the descriptive aspects of economic theory, that is, both welfare economics and the ability of economic models to explain behavior. Sen summarizes his view thus: "The object is to understand, explain and predict human behaviour in a way such that economic relationships can be fruitfully studied and used for description, prognosis, and policy. The jettisoning of all motivations and valuations other than the extremely narrow one of self-interest is hard to justify on grounds of predictive usefulness, and it also seems to have rather dubious empirical support." (Sen 1987: 79)

5. Agency as norm responsive behavior

Although it is important to recognize the multiplicity of motivations agents have, the way Sen draws the distinction between the agency-aspect and the well-being aspect of persons does not help us to appreciate how his theory contrasts with the standard philosophical theory of agency. According to Sen, committed behavior is an expression of the agency aspect of persons, but some of what Sen calls behavior motivated by "commitment" is not self-goal directed and thus not an expression of agency on the standard view. The well-being aspect of persons is to be *contrasted with* the agency aspect and thus it is not an expression of agency. Persons' understanding of their well-being can motivate them to act, but well-being is not their only motivation. As Sen has argued, well-being understood in the narrow sense of self-interest captures only an aspect of the welfare of agents, and not necessarily their goals, let alone motivations for choice. Sen's view that commitment can motivate action without its being the agent's own goal contrasts with agency as understood on the standard model, and yet this contrast is not captured in the agency/well-being distinction.

I believe that there is a more basic "essential duality" in Sen's characterization of action based on commitment as opposed to self-interest, which allows both types of behavior to be seen as expressions of agency. I will call this the duality of autonomy-agency vs. identity-agency. While autonomy-agency is agent-goal directed, identity-

agency is other-goal directed. Explaining Anil's choosing the mango by referring to his wish to impress his girlfriend with his commitment to cook authentic Indian food for the dorm fundraising dinner is an expression of autonomy-agency. Explaining Anil's choosing the mango by referring to his habit of cooking the authentic food of his homeland is an expression of identity-agency. Autonomy-agency is explained as self-goal directed (in pursuit of self-welfare goals or goal-modifying commitments); identity-agency is explained by goal-displaced commitments as motivating factors.

There are two points to emphasize, and to keep distinct, about the importance of commitment to the explanation of human behavior. The first is that commitments are crucial to understanding the agency aspect of persons, and the second is that seeing persons as acting on commitments is indispensable to the explanation of some behavior. The first point is one about autonomy-agency, which implies acting on reasons that are one's own. If the behavior is to be seen as action expressing autonomous agency, then the goal has to be the agent's own.[15] Commitments that express autonomous agency either involve self-goal choice directly or indirectly through modifying the original self-goal and becoming a new self-goal on which the agent acts. But Sen also recognizes that commitments can motivate by displacing rather than modifying an agent's goals, such as when unconsciously following social norms or expressing group identity.[16] Sen writes, "rejection of self-goal choice reflects a type of commitment that is not able to be captured by the broadening of the goals to be pursued. It calls for behavior norms that depart from the pursuit of goals in certain systematic ways. Such norms can be analyzed in terms of a sense of 'identity' generated in a community…" (Sen 1985a: 219)

Such behavior is clearly common and important to include in explanatory models of human behavior. At the same time goal-displacing commitments seem to deny agency, when agency is understood as the agent acting on his own goal. So although we seem to need both kinds of agency in our explanations of behavior, identity-agency looks to be an

---

[15] I see autonomy as minimally requiring the agent's authentic identification with that motivation.

[16] (Bicchieri 2006: ch. 4) has an illuminating discussion of group identity and social norms along these lines.

oxymoron. Goal-displacing commitment based explanations depart from intentional rational choice explanations, and thus from the standard descriptive model of agency. Either they involve 'as if' explanations where agents sublimate their own goals for the sake of taking on a communal identity where the causal mechanism being posited is the communal goal that motivates the agent, or they involve behavior that is conditioned by evolved behavioral regularities that also override individual goals. Either way they are explanations that take the mechanism of behavior to be external to the individual agent.

This implies, I think, that we need to posit a broader theory of human agency, in which human agency can be expressed as fundamentally social or biological in origin, rather than intentional. In this theory of agency, to be an agent is to act responsively in a normative framework. Formulating one's conception of the good is one kind of response to a normative framework. Thus, the standard theory of agency as involving formulation of an end and acting in light of it is one type of agency on this broader view. Another way to act responsively in a normative framework is to act within normative constraints that one has internalized, but that one may not be consciously attending to. On this broader theory of agency, what makes behavior count as an expression of human agency is the fact that it is norm governed, though the norms need not be intentionally acted upon. Identity-agency, on this view of agency, is not an oxymoron.

6. Tacit commitment and identity

Goal-displacing commitment as a behavioral motivation remains puzzling because it is unclear to what degree we should see it as a species of action explanation. On the one hand it seems that a commitment can explain action only if it plays a role in the motivation of the action, and that seems to require it do so as a reason for action. On the other hand, it seems that as soon as it becomes a reason for an agent's action it thereby becomes *her* reason for action, and thus a species of goal-modification, not goal-displacement. I want to suggest that there is a type of goal-displacing commitment that can be seen to fit between the horns of this dilemma. I will call this "tacit commitment."

Tacit commitments, I propose, regulate behavior without being made conscious and explicit. Such behaviors include cliquish behavior that isn't recognized as such, or conforming to background norms that one never questions. Why does Anil always cook

Indian food? Why do we wear formal suits in the summer in Kansas, where it is 100 degrees (Fahrenheit) and humid? What accounts for product loyalty when the product is not clearly superior to others? Sure, sometimes we conform out of a conscious wish to do so, but sometimes we go along with such crowd behavior without examining it. What makes these behaviors 'commitments' in Sen's sense is that they form a basis for group identity. Some tacit commitments may be less subtle and more socially complex, such as the norms of social distancing that we master as children, or the racial and ethnic prejudices that are instilled in many communities. Yet, once such commitments are recognized as guiding one's behavior, they may or may not be embraced for future behavior. Of course, once embraced explicitly, such commitments are no longer tacit and behavior explained by reference to them is now explained in the standard intentional rational choice manner (though as neither self-centered welfare nor self-welfare goal).

By tacit commitments I mean to refer to behavioral causes that are even more tacit than conventional rules. Tacit commitment is a kind of external motivation; it is not a commitment to a set of norms or constraints chosen from among other possible sets. Rather, an agent acting on a tacit commitment takes the norms or constraints as behavioral guides and others interpret that behavior as indicating something about the identity of the individual.

Sen's 2006 book, *Identity and Violence*, explores ways that persons take on identities that they seem to feel destined to embrace. He argues that we ought to see ourselves and each other as constituted by many different identities. And we should embrace the freedom to choose our identities and to choose not to identify with those group commitments that lead to hatred, prejudice, and violence. I wholeheartedly agree with his view. However, as Sen well recognizes, there are also forms of identity that are ascribed, non-voluntary, and non-intentional. This is why we feel destined to embrace identities even when we do not feel that by doing so we improve our well-being or achieve any of our own goals. Oppressed persons often feel destined to embrace their identity as an oppressed person because they are inevitable and inescapable, not because

the identity brings them well-being, pleasure, or dignity.[17] These identities are constituted, I suggest, by tacit commitments that motivate behavior without involving deliberation or invoking questioning on the part of the agent. They are externally created and imposed, but internalized in the behavioral patterns of individuals. The identity that they create may be externally imposed as an ascribed identity. But they are enacted by an agent (exercising identity-agency) who takes them for granted and behaves according to them. Tacit commitments can be made explicit, questioned, and rejected. But insofar as they tacitly guide our behavior they constitute, at least in part, our ascribable group identities.

Can tacit commitment be accommodated within an explanatory theory of action as maximization of an objective function, a self-goal? I believe that it can. Because tacit commitments guiding persons can be made explicit by the observing economist, they can be modeled as goals that the agent has, since he acts as if he is making his behavior correspond to them. Yet, being goals of the group that are unacknowledged by the agent, they are not realistically assumed to be the agent's own goals.[18] In this way we can see tacit commitment as playing both a motivating role in the behavior of the agent, and yet still not representing an actual, explicit, intentional commitment of the agent. The commitment is therefore not the self-goal choice of the agent, but rather a community goal.[19] But given a goal that the agent is pursuing, the maximizing model of behavior can

---

[17] See, for example, discussion of the complicated preferences of the Sufi Pirzada women in (Narayan 2002), or the adaptive preferences of the women of El Pital in (Khader 2011: ch. 1).

[18] Tacit commitments can also be agents' own goals even if they are not aware of their own goals. Ordinary habitual behavior can be automatic and unconscious, but could still reflect the agent's own goals, because it was originally acquired in pursuit of the agent's conscious, intentional goals and still serves those ends. I thank Elizabeth Anderson for helping me to clarify this point.

[19] For a different account of the connection of goal-displacing commitments to community goals that involves intention, see (Schmid 2007).

be invoked. Furthermore, we can recognize acting on tacit commitments as an expression of identity-agency.

It is important to recognize, categorize, and theorize tacit commitments as they actually exist, as unrecognized by the agents who are nonetheless motivated to behave by them. Yet this requires a kind of behavioral science that Sen himself does not engage in. This kind of motivation will be relative to context and situation, and discerning the social norms in play in any given situation will require thick descriptions of local culture. As Sen has long recognized, humans are social beings and choices are always social acts. (Sen 1973: 252-3) Social psychologists and anthropologists do investigate these kinds of motivations, and so it will be important to use their results to impute the right tacit commitments in economic models that must make use of them.

One objection that might be raised to such multidisciplinary explanatory pluralism is that there is no overarching theory to specify which explanatory schema is to be used in a given instance. When should we ascribe the tacit 'as if' commitment to explain a behavior, as opposed to seeing the behavior as guided by an internalized social norm, or as guided by reputation seeking for long-run self interest? By allowing such a wide variety of models and goals we risk the charge of adhocness in any particular explanation. Although this charge raises a caution for explanatory pluralism, I do not think it overrides the benefits of seeing behavior as guided and motivated by a variety of forces internal and external to the agent. While some of the predictive value of the theory is sacrificed, much is gained in the ability to describe and evaluate behavior with this explanatory variety. (Sen 1980, 1986a.) As a philosophical theory of agency, however, this charge of adhocness does not apply.

7. Conclusion

My interpretation of Sen's use of goal-displacing commitment to explain behavior has the virtue of allowing Sen to explain a wider variety of behavior and yet not making a conceptual mistake about the nature of agency, as Pettit charges. Instead we can see Sen as giving an alternative foundation to the theory of human agency; agency is fundamentally norm-governed behavior and consciously recognized norms comprise only some of those norms. It also allows the use of the maximizing model of rationality, which

conforms to his response to Anderson. Although the maximizing model of rationality can be invoked with tacit commitment explanations, the agent is seen as acting autonomously only if he or she embraces the rationalization of the model.[20] I argued earlier that Anderson was correct to point out that Sen has not shown why it is rational to act as if one has preferences or goals that one does not in fact take as one's own. In some cases, as I argued, agents are rational to act as if they have preferences other than the ones they in fact have because by doing so they better achieve the satisfaction of preferences they do have. But Sen need not show that tacit commitment is itself a rational motivation of the agent who acts when she acts. For the agent acting out of tacit commitment, the motivation is not intentional. The commitment need only be rationalizable when considered from the perspective of the observer, or from the perspective of the agent once it is made explicit. There is a danger of rationalizing too much behavior; there is always the possibility that behavior is irrational or self-deceptive, after all. The key here to formulating good social science about irrational or non-rational behavior is to find the systematic regularities of social situations or neurotic or other internal causes of behavior that justify resorting to such explanations in favor of rationalizing ones. There is of course a lot to say about such modeling decisions, but that would take us too far afield.

In this paper I have tried to interpret Sen's account of behavioral motivation and his critique of standard economic models of behavior. For Sen, the purpose of economic science is to "to understand, explain and predict human behaviour in a way such that economic relationships can be fruitfully studied and used for description, prognosis, and policy." (Sen 1987: 79) I take this to be a summary of his pluralism and pragmatism about social science models. This can equally well be seen from the fact that he recognizes and discusses a variety of different and competing models of behavior, suggesting that different ones are appropriate for different situations. If we take economics to help us by way of describing and prescribing motivations, then we can see that each of the explanatory schemas that Sen discusses allows for a different possible

[20] There may be additional requirements for acting autonomously; embracing the commitment or norm motivating one's action is a necessary, not sufficient requirement for autonomy.

source of motivation: self-interest, sympathy, social norms, conventional rules, group oriented goal modifying commitments, and goal-displacing commitments. On some of these schemas, we can model behavior as intentional and rational via the standard maximizing model, while on some others behavior is non-intentional, though rationalizable. Sen allows that there are some behaviors best understood as primarily individually oriented and internally motivated, while others are best understood as motivated externally by group norms. These forms of motivated behavior are best seen as expressions of human agency broadly conceived, types of which I have termed autonomy-agency and identity-agency.

In his paper "Prediction and economic theory" Sen refers to two aspects of complexity that make prediction in economics difficult, one is the choice problem, which is just the problem of the many different kinds of factors – "social, political, psychological, biological, and other factors" – that influence behavior, and the other is the interaction problem, which arises from the interactions of many individuals whose behavior is influenced by so many factors, as well as "different values, objectives, motivations, expectations, endowments, rights, means, and circumstances dealing with each other in a wide variety of institutional settings." (Sen 1986a: 5) These aspects reflect both the motivational pluralism Sen acknowledges in his critique of standard economic models of behavior, and his recognition of the normative situatedness of human behavior.

On Sen's account, explanation requires social contextualization whether we model behavior as internally motivated by self-interest or externally motivated by social norms. He argues that we cannot get sufficiently accurate predictions or realistic explanations by overlooking commitments. "The jettisoning of all motivations and valuations other than the extremely narrow one of self-interest is hard to justify on grounds of predictive usefulness, and it also seems to have rather dubious empirical support." (Sen 1987: 79) But commitments are sometimes a group interest or social norm, which is necessarily contextual. Although Sen recognizes the need for context, there is no reason to think that this cannot be pursued scientifically by social psychologists and anthropologists. Yet, as he notes, the work of social and behavioral sciences is fundamentally different from that of physical scientists in needing to take intentions into account as additional kinds of causal mechanisms.

Given the normative situatedness of human behavior, values and facts cannot be separated entirely from each other. Economics must always take norms into account as among the facts to be explained, and is itself a normative project when its models prescribe actions for individuals and social policy. Sen has been a forceful proponent of recognizing and embracing the interconnections of ethics and economics, continuing a tradition that he often traces to Adam Smith. Economics stands to be a better explanatory social science by enhancing its understanding and appreciation of the sources of motivation beyond private self-interest.

References

Anderson, Elizabeth, 2001, "Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Economics and Philosophy," *Economics and Philosophy*, 17(2001): 21-38.

Bicchieri, Cristina, 2006, *The Grammar of Society: The Nature and Dynamics of Social Norms*, New York: Cambridge University Press.

Binmore, Ken, 2005, *Natural Justice*, New York: Oxford University Press.

Bossert, Walter and Kotaro Suzumura, 2009, "External Norms and Rationality of Choice," *Economics and Philosophy*, 25(2009): 139-152.

Cudd, Ann E., 2006, *Analyzing Oppression*, New York: Oxford University Press.

Dimock, Susan, 1999, "Defending NonTuism," *Canadian Journal of Philosophy*, 29(1999): 251-274.

Dohngde, Shatakshee and Pattanaik, Prasanta K., 2010, "Preference, Choice, and Rationality: Amartya Sen's Critique of the Theory of Rational Choice in Economics," in *Amartya Sen*, pp. 13-39.

Dowding, Keith, 2002, "Revealed Preference and External Reference," *Rationality and Society*, 14(2002): 259-284.

Elster, Jon, 2007, *Explaining Social Science: More Nuts and Bolts for the Social Sciences*, Cambridge: Cambridge University Press.

Friedman, Milton, 1966, "The Methodology of Positive Economics," in *Essays In Positive Economics,* Chicago: Univ. of Chicago Press, pp. 3-43.

Gauthier, David, 1986, *Morals by Agreement*, New York: Oxford University Press.

_____, 1996, “Commitment and Choice: An Essay on the Rationality of Plans,” in *Ethics, Rationality, and Economic Behaviour*, Farina, Hahn, and Vannucci, eds., Oxford: Oxford University Press, 1996, pp. 217-243.

Hausman, Daniel M., 2005, “Sympathy, Commitment, and Preference,” *Economics and Philosophy*, 21, pp. 33-50.

Khader, Serene, *Adaptive Preferences and Women’s Empowerment*, New York: Oxford University Press, 2011.

Kreps, David M., Milgrom, Paul, Roberts, John, Wilson, Robert, 1982. "Rational cooperation in the finitely repeated prisoners' dilemma," *Journal of Economic Theory*, 27(2): 245-252.

Morris, Christopher, ed., 2010, *Amartya Sen*, Cambridge: Cambridge University Press.

_____, 2010, “Ethics and Economics,” in *Amartya Sen*, pp. 40-59.

Narayan, Uma, 2002. “Minds of Their Own: Choices, Autonomy, Cultural Practices, and Other Women,” in *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, ed. Louise M. Antony and Charlotte E. Witt, Westview Press, 2nd edition, 418-432.

Oldenquist, Andrew, 1982, “Loyalties,” *Journal of Philosophy*, Vol. 79, No. 4 (April), pp. 173-193.

Pettit, Philip, 2005, “Construing Sen on Commitment,” *Economics and Philosophy*, 21: 15-32.

Schmid, Hans Bernhard, 2007, “Beyond Self-Goal Choice: Amartya Sen’s Analysis of the Structure of Commitment and the Role of Shared Desires,” in Rationality and Commitment, Peter and Schmid, eds., Oxford: Oxford University Press, pp. 211-226.

Sen, Amartya, 1974, “Choice, Orderings and Morality,” in *Practical Reason*, Körner, ed., Oxford: Blackwell Press, pp.54-67.

_____, 1973, “Behavior and the Concept of Preference,” *Economica*, (August), pp. 241-259.

_____, 1977, “Rational Fools: A Critique of the Behavioral Assumptions of Economic Theory,” reprinted in Hahn and Hollis, eds., *Philosophy and Economic Theory*, New York: Oxford, 1979, pp. 87-109.

_____, 1980, “Description as Choice,” *Oxford Economic Papers*, New Series, Vol. 32, No. 3 (Nov.): 353-369.

_____, 1982, "Rights and Agency," *Philosophy & Public Affairs*, Vol. 11, No. 1, 3-39.

_____, 1985a, "Goals, Commitment, and Identity," *Journal of Law, Economics, and Organization*, 1(Fall, 1985), reprinted in Sen, *Rationality and Freedom*, pp. 206-224.

_____, 1985b, "Well-Being, Agency, and Freedom: The Dewey Lectures 1984," *Journal of Philosophy*, Vol. 82, No. 4 (April): 169-221.

_____, 1986a, "Prediction and Economic Theory," *Proceedings of the Royal Society London*, A (407): 3-23.

_____, 1986b, "Rationality, Interest, and Identity," in *Development, democracy, and the art of trespassing : essays in honor of Albert O. Hirschman*, Alejandro Foxley, Michael S. McPherson, and Guillermo O'Donnell, eds., Notre Dame: University of Notre Dame Press, pp.343-353.

_____, 1987, *On Ethics and Economics*, Oxford: Basil Blackwell.

_____, 1993, "Internal Consistency of Choice," *Econometrica*, 61(May, 1993): 495-521.

_____, 1997, "Maximization and the Act of Choice" *Econometrica,* 65(July, 1997): 745-79, reprinted in *Rationality and Freedom*, pp. 158-205.

_____, 1999, *Development as Freedom*, New York: Random House.

_____, 2002, *Rationality and Freedom*, Cambridge: Harvard University Press.

_____, 2005, "Why exactly is Commitment Important for Rationality?" *Economics and Philosophy*, 21(2005): 5-14.

_____, 2006, *Identity and Violence: The illusion of destinty*, New York: W.W. Norton.

_____, 2009, *The Idea of Justice*, Cambridge: Harvard University Press.