RMM Vol. 4, 2013, 157–172
http://www.rmm-journal.de/

Christoph Hanisch

Negative Goals and Identity: Revisiting Sen's Critique of *Homo Economicus*[*]

Abstract:
Sen's critique of the *homo economicus* conception of choice asserts that agents who 'displace' their goals, and instead choose on the basis of others', are not therefore irrational. I first defend Sen against the objection that violations of "self-goal choice" undermine coherent deliberation. My critique of Sen then introduces the notion of 'negative goals' and shows that the process of adopting others' aims remains constrained by those 'goals' that determine the spectrum of actions that an agent considers permissible. Only on rare occasions are we pushed to violate even these negative goals that play a central role for our identities.

Keywords: Rational choice theory, *homo economicus*, practical identity, self-goal choice, commitment.

1. Introduction

In a number of publications Amartya Sen has presented a critique of orthodox rational choice theory. In "Rational Fools" Sen (1977) introduces the notion of "commitment" in order to undermine the conception of *homo economicus* that underlies much of contemporary economics and the account of human behavior employed there. Human agents are, according to Sen's alternative conception, not exclusively concerned with maximizing their own narrowly conceived welfare; our identities are more richly structured and the possibility of committed choice and action shows that much of economic theorizing rests on an inadequate account of rationality.

In this paper I discuss a three-fold distinction that Sen (1985) introduces in the essay "Goals, Commitment, and Identity" and that further develops his original criticism of the *homo economicus* paradigm (see also Sen 2002, 33–37). I focus on the third dimension of narrow self-interest that Sen attacks there, namely "self-goal choice". Whereas the possibility of choices and actions that run contrary to the first two aspects of narrowly defined self-interest (i.e., "self-centered

[*] I want to thank Michael Weber and the members of Fred Miller's research group at Bowling Green State University for very valuable comments on earlier versions. At later stages the paper profited greatly from the comments provided by two anonymous referees for this journal. Research for this paper was funded by the ERC Advanced Grant "Distortions of Normativity".

welfare" and "self-welfare goal") are now widely accepted as necessary extensions of traditional accounts of rational choice, critics have been much more reluctant to accept Sen's claim that commitment enables rational agents to choose on the basis of others' goals. Sen's formulation of how agents can violate the feature of "self-goal choice", without thereby counting as irrational, is ambiguous. It is not clear what it means for an agent not to choose on "her own goals". This essay is to a large extent an attempt to clarify this ambiguity. The plausibility of Sen's critique of rational choice theory is dependent on what it means to choose (and not to choose) on the basis of one's own goals and, *a fortiori*, what it means to choose on the basis of other agents' goals *without* thereby making these goals one's own. Do committed agents literally take-in others' goals (and the related intentional states) or does Sen have something far less controversial in mind, when he considers cases of practical deliberation and action that contradict the feature of self-goal choice?

Philip Pettit (2005), interpreting Sen as making the more ambitious claim (committed agents take-in others' goals), argues that conformity with self-goal choice is constitutive of being a conscious agent who intentionally chooses and acts. Hence, according to Pettit's interpretation, Sen's view that committed *agents* are able to transcend their own goals is unrealistic, even incoherent. Violations of self-goal choice and acts of "goal-displacing commitment" conflict with common sense and with the minimal conception of rational choice theory (defended by Pettit), according to which an agent necessarily chooses *exclusively* on the basis of her own, currently held, goals.

Whereas Pettit's conclusion, *viz*. that rational agents cannot violate the feature of self-goal choice, is basically correct (I discuss an important exception in *section 5*), he arrives at it for the wrong reasons. His argument rests on adopting the first, ambitious and controversial, answer to the question of what Sen's ambiguous definition of self-goal choice means. Pettit attributes the psychologically implausible view to Sen, according to which committed agents literally absorb others' goals and intentions. Sen can be defended against this charge by means of introducing the distinction between other agents' intentions and their goals' content. The content of another person's goals are the particular reasons for choice and action that are reflectively endorsed by the individual in question when she is committed to promote the other person's goals. The other person's intentions, on the other hand, cannot be "absorbed" by her.

This defense of Sen notwithstanding, I formulate my own worries concerning his view that commitment sometimes results in choice that violates self-goal choice and that an agent sometimes adopts the goals of others "without their taking the form of goals that the person can be seen as pursuing himself" (1985, 347). I claim that transcending self-goal choice, while not psychologically impossible, constitutes a radical challenge to our identities as unified moral agents. In order to establish this claim I introduce the notion of "negative goals". These goals function as normative constraints, limiting the kinds of foreign goals that we are willing to adopt in our choices. The debate about goals and preferences in economic theory (and the Sen-Pettit exchange is no exception) has been dom-

inated by a concern with positive goals, i.e., states of affairs that an agent actively strives to bring about and realize. My claim is that paying attention to negative goals is necessary to flesh out Sen's notion of commitment. In so far as agents have stable moral identities their choices and actions always reflect *their* goals, even if the latter sometimes play a merely negative role in establishing the boundaries of a spectrum of permissibility within which they adopt others' positive goals and the reasons that are provided by them. Pettit is right when he claims that we (almost) always choose on the basis of *our* own goals; however, the best explanation for this correct observation is one related to our stable agential identities and not a psychological one.

I conclude by discussing one objection to my account of negative goals. In rare cases violations of self-goal choice (even choices transgressing one's negative goals) *are* possible. Bernard Williams' (1973, 98–100) example of Jim and the Indians illustrates my reply to this objection according to which an agent's negative goals (what "she cannot see her*self* doing") are partly constitutive of her identity as a unified agent. Transcending all of one's goals by transgressing these constraints is tantamount to a reconstitution of an agent's identity. In so far as Sen understands commitment as the possibility of violating these negative constraints I draw attention to the *prima facie* significant personal costs and strains that are involved in such violations. Sen overlooks these non-negligible costs when he seems to insist that the possibility of goal-displacing commitment is among the most rational and noblest expressions of human selflessness. Without denying that this selflessness may sometimes be required on moral grounds, I insist that genuine reconstitutions of personal identities always have a price. It is no coincidence that relocating one's negative goals (and thereby changing one's identity) is regarded as a truly *self*less act.

2. Sen on Self-Goal Choice

The central claim of Sen's paper "Rational Fools" is to disassociate an agent's welfare-achievement from her choice behavior. Sen's project is to show that the conception of *homo economicus* manifests an overly narrow account of how human beings choose and behave. We are, according to Sen, not always and exclusively concerned with the maximization of our narrowly conceived self-interest. Agents are receptive to influences of both sympathetic sentiments towards others and to commitment, only the latter posing the genuine challenge to the *homo economicus* paradigm under discussion: "The characteristic of commitment with which I am most concerned here is the fact that it drives a wedge between personal choice and personal welfare, and much of traditional economic theory relies on the identity of the two." (1977, 329)

Sympathy, which Sen (1977, 326) discusses first, "corresponds to the case in which the concern for others directly affects one's own welfare", e.g., seeing the suffering of others affecting one's own well-being. Sympathy can easily be incorporated into traditional rational choice theory and the *homo economicus*

paradigm though. In fact, as Sen (2005, 7) acknowledges, Gary Becker (1976; 1996) dedicated a lot of his work to the question of how welfare-maximization takes into consideration the fact that we sympathize with others and their welfare-achievement. According to Becker, orthodox economic theory merely needs to broaden its conception of individual welfare in a way that allows sympathy (and antipathy) to get incorporated into the individual utility calculus. This egoistic extension does not constitute a departure from the self-interested *homo economicus*.

Things are different with respect to commitment: "commitment does involve, in a very real sense, counterpreferential choice, destroying the crucial assumption that a chosen alternative must be better than (or at least as good as) the others for the person choosing it, and this would certainly require that [economic] models be formulated in an essentially different way." (Sen 1977, 328) Commitment is not necessarily a moral disposition to choose and to act (it can be based on custom and habit or on highly particularistic affiliations). The aforementioned wedge between choice and welfare consists in Sen's claim that human agents very often do not choose and act in a way that suggests maximization of their own welfare. When, for example, agents choose to promote a cause like global social justice, without the successful decrease in suffering affecting their own welfare, then this is a case of choice based on commitment. This kind of choice constitutes a departure from traditional (Beckerian) rational choice theories which do not admit that such choices, which are unrelated or even contrary to one's own welfare, count as rational ones.

In "Goals, Commitment, and Identity" Sen (1985, 346–349) refines and radicalizes his attack on the *homo economics* paradigm by relocating the conceptual wedge driven by commitment. He now sees the wedge as separating an agent's *choices*, on the one hand, from her *goals*, on the other (rather than, as in "Rational Fools", between choice and welfare). More precisely, in this essay Sen introduces three aspects of "privateness", i.e., three distinct elements of self-interest that underlie the economic view of agency that he attacks. Sen's (2002, 33–34) most recent re-statement defines these "three different ways in which the self may be central to one's self-interested preferences and choices" as follows:

> "*Self-centered welfare*: A person's welfare depends only on her own consumption and other features of the richness of her life (without any sympathy or antipathy towards others, and without procedural concern).
> *Self-welfare goal*: A person's only goal is to maximize her own welfare.
> *Self-goal choice*: A person's choices must be based entirely on the pursuit of her own goals."

According to Sen, traditional economic conceptions of choice and behavior group together these three elements under the heading of "self-interest maximization". The picture presented by Sen brings to attention the diversity of assumptions that underlie the framework of self-interested rationality. The three aspects are

independent of each other, e.g., allowing the welfare of distant others to enter one's set of goals constitutes a violation of the second aspect (self-welfare goal) but does not necessarily say anything about whether or not the other two aspects of self-interest are transcended (see, e.g., Sen 1985, 347; Sen 2005, 6). The same is true of the other two ways in which the self is central to the *homo economicus*' self-centered preferences and choices. Recently, Sen (2002, 36) supplements these three with a fourth, more abstract, feature of the self relevant for choice, namely "the one that is able to do self-scrutiny and reasoning". It is this fourth feature that is responsible for the possibility of committed choice and action. It is also the self's feature that rational choice theory ignores and we will examine it in more detail below when we moderate the Sen-Pettit exchange.

Becker's rational choice theory, for example, merely recognizes the untenability of the first aspect of self-interest as a necessary condition for rational choice. He incorporates the direct impact that others' welfare has on our own, thereby making room for rationally permissible violations of the requirement of self-centered welfare. However, the other two ways in which narrow self-interest can be transcended (by means of violating self-welfare goal and self-goal choice respectively) are more difficult to incorporate into traditional accounts of rational choice. The wedge driven by commitment between welfare and choice has already been mentioned and the challenge to Becker there consists in tolerating the violation of the rational choice dogma of self-welfare goal. Here I focus on Sen's even more substantial and ambitious critique of the *homo economicus* paradigm, namely the claim that we do not always choose on the basis of our own goals. The wedge at issue now consists in how some commitments disassociate one's choices from the set of goals one regards as pursuing at the moment of choice. That the possibility of such a disassociation is not merely a fact about us but an essential element of human rationality is the central claim of Sen's recent work.

3. Pettit's Critique of Goal-Displacing Commitment

A violation of self-goal choice takes place when agents take into consideration other agents' goals, either as constraints when pursuing their own goals or, even stronger, when their choices are based on goals that are not incorporated into their own goal-sets at the moment of choice. Philip Pettit (2005, 18) identifies two ways in which Sen lets commitment account for the violation of self-goal choice. Sen distinguishes between "goal-modifying commitment" and "goal-displacing commitment".[1] Both types of commitment call into question self-goal choice, but only the second one is incompatible with rational choice theory in the "minimal understanding" that Pettit (2005, 16) defends. He says: "To believe in the minimal version of the [rational choice] theory is simply to believe

[1] As Pettit notes, Sen does not use these two concepts but he is right to present the distinction between goal-modifying and goal-displacing commitment as being implicit in Sen's discussion.

that one's choices are based on one's own goals; it is to believe in what he [Sen] calls 'self-goal choice'."[2]

In fact, Pettit (2005, 18) claims, goal-modifying commitment is not really a violation of self-goal choice and hence remains compatible with (minimal) rational choice theory: "I will continue to promote my modified goals—I will maximize in the familiar pattern—though the goals I come to serve will no longer be the goals of self-interest, even enlarged self-interest." We modify *our* own goals, plans, and pursuits from time to time. I may be engaged in the process of learning to play the piano with the ultimate goal of doing so like Alfred Brendel, pursue this goal for a couple of years but eventually realize that my goal has been too ambitious. In the face of realizing my limited potential I modify my goals and accept that my career as a pianist may not get me any farther than my hometown's semi-professional orchestra. Similarly, we are sensitive to how the pursuit of our goals conflicts with other people's goals. A person may modify his goal of pursuing an eccentric but economically unpredictable artistic career on learning that his partner is pregnant. His prospective child's interests and goals require him to settle with a more commercially oriented way of pursuing his artistic desires such as accepting a less exciting job as a commercial architect.

In contrast to goal-modification the problem with goal-displacing commitment lies, according to Pettit, in Sen's talk about violating self-goal choice by fully adopting other people's goals. Goal-modifying commitment that takes into consideration other people's goals (like in the examples just presented) is unproblematic because we continue to modify our own goals when we see that their pursuit negatively impacts on other people's lives. However, *replacing* one's own goals with the goals of others is one step too many taken by Sen. According to Pettit, Sen's defense of transcending self-goal choice amounts to accommodating as intelligible the choices of agents who fully internalize other people's goals. Such 'agents' choose on the basis of goals that no longer reflect goals of their own at all. Pettit (2005, 19) summarizes:

> "But Sen [. . .] maintains that people may be committed to others in such a way that they no longer act [. . .] on their own goals; 'the pursuit of private goals may well be compromised by the consideration of the goals of others'. People may become the executors of a goal-system that outruns the private goals that they endorse in their own name: a goal-system that makes place for the goals of others or for the goals of groupings in which people cooperate with others."

[2] Pettit (2005, 16–17) distinguishes between a minimal conception of rational choice theory (the one he defends against Sen and that is compatible with a common-sense understanding of goal-modification), on the one hand, and a more substantial version of rational choice theory à la Becker that insists on self-welfare goal, on the other. By defending this distinction Pettit admits that Sen's claims about the violation of the second feature of self-interest (self-welfare goal) is, contrary to Becker, plausible. However, as will be spelled out in the text, it does not follow that this minimal conception of rational choice theory (claiming that we *always* and *necessarily* choose on the basis of some of *our own* goals) is threatened even when such violations of self-welfare goal are, *pace* Becker, allowed, or so Pettit claims.

Pettit (2005, 19) finds this view "highly implausible" since goal-displacing commitment amounts to no longer choosing and acting on goals that one endorses from one's own deliberative standpoint. He claims that Sen is violating fundamental intuitions concerning intentionality, underlying common sense psychology and the minimal conception of rational choice theory under attack. In fact, Pettit claims, Sen's defense of the rationality of violating self-goal choice (by means of goal-displacing commitment) is on the verge of being a conceptual impossibility with respect to intentionality, choice, and action. After all, how should we conceive of an agent who acts on the basis of goals that are not at all hers? Pettit reminds us that intentionality requires that an agent incorporates (other agents') goals into her own goal-set in order to choose and act *on them*. Only if this incorporation is fully performed is the agent in question in a position to pursue others' goals (which are then no longer others' goals but have become hers anyway). Sen is violating this common-sense view of intentionality by his advocacy of rationally violating self-goal choice in the form of goal-displacing (as opposed to goal-modifying) commitment. In Pettit's (2005, 21) words: "To imagine an action that is not controlled by a goal of the agent, by the lights of this approach, will be like trying to imagine the grin of the Cheshire cat in the absence of the cat itself. Let the agent not have a goal and it becomes entirely obscure how the agent can be said to act; to act, or at least to act intentionally, is to act with a view to realizing a goal."

Pettit's objection rests on a peculiar reading of Sen's passages on violating self-goal choice and interpreted in context, Sen's argument can be rendered more plausible and be defended (within limits) against Pettit's criticisms. In order to defend Sen one must keep in mind that in addition to the three features of the self, as incorporated in orthodox rational choice theory, Sen (2002, 34-36) presents a fourth element, namely the one making commitment possible by means of reflective scrutiny and practical reasoning. This feature therefore takes on the role of a kind of 'corrective' element of the rational agent's capacities. Bringing it to our attention is ultimately the key to rectifying the major misunderstandings underlying classical rational choice theory. Sen stresses that,

> "[a] person is not only an entity that can enjoy one's own consumption, experience and appreciate one's welfare, and have one's goals, but also an entity that can examine one's values and objectives and choose in the light of those values and objectives. Our choices need not relentlessly follow our experiences of consumption or welfare, or simply translate perceived goals into action."

Taking into consideration this fourth dimension does not merely account for the possibility of cooperative behavior in prisoner's dilemma cases—a problem Sen is concerned with in his writings on how commitment resolves collective action problems (1977, 340–341; 1985, 342–346; see also 1982, part I). In addition, our nature as scrutinizing and reflecting agents helps to make sense of *what* is happening in cases of goal-displacing commitment and when we 'replace' our

goals with those of other individuals or with collective goals (such as upholding just social institutions) and take these goals as the basis of our choices. Our capacity for practical reasoning is responsible for the feature of transcending self-goal choice that Pettit finds so troublesome and counterintuitive, namely the way in which others' goals exert "influences [that] affect the person's choice without their taking the form of goals that the person can be seen as pursuing himself" (Sen 2002, 214).

According to my interpretation, Sen's talk about displacing one's goals with those pursued by other agents works much like the case of goal-displacement with respect to our own goals over time.[3] In both cases we need to distinguish between a goal understood as the feature of an individual's intentional stance, on the one hand, and a goal understood as a cognitive item providing a specific content of actual deliberation (i.e., particular reasons for choice and action), on the other. In the case of goal-displacement I am not putting another person's intentions in the place of my own in the literal sense that Pettit attributes to Sen. However, and now the fourth feature of the self mentioned above becomes relevant, a deliberating and choosing individual is very well capable of integrating other agents' reason-providing goal-features into her choices in a fairly thorough sense. According to Sen (2002, 4), rationality is to be understood as "the discipline of subjecting one's choices [...] to reasoned scrutiny". It "is seen here [...] as the need to subject one's choices to the demands of reason". Whether or not an agent regards the reasons constituted by the goals that others pursue as strong enough and worthy of her endorsement depends, according to Sen (1985, 348), on the individual's self understanding: "[I]n arriving at goals, a person's sense of identity may well be quite central. And, perhaps most important in the context of the present discussion, the pursuit of private goals may well be compromised by the consideration of the goals of others in the group with whom the person has some sense of identity."[4] Our capacity for practical reasoning enables us (though it obviously does not necessitate us) to choose paths of actions and policies on the basis of considerations that we regard as trumping concerns, while at the same time, we do not explicitly endorse those very goals that others pursue and that are the ultimate source of these concerns.

The eccentric artist then might not merely modify his goals and choice-behavior in order not to negatively affect his child's interests and goals. Rather, and more radically, the prospective father must be seen as replacing his current goal (qua basis and reservoir of reasoned choice) of being an artist entirely with the goal of being an enthusiastic full-time father, co-pursuing his child's goals *by means of* taking the related reasons for choice and action as normative requirements. It is important to note that this latter case of goal-displacing commit-

[3] Such a case would be giving up playing the piano at all, and becoming a novelist instead, upon realizing that one's capacities for the former activity are too limited.

[4] Recently Sen (1999; 2004) has presented an account of agents' "social identities" and its relationship to commitment. This account is critical of communitarian views. The latter tend to deny the possibility of Sen's fourth (abstract) element of the self discussed above.

ment comes very close to a significant change in the artist's personal and practical identity and I will discuss this issue below.

We can, I submit, regard such a case of "putting others' goals in place of one's own" as an instance of genuine goal-*displacing* commitment that does not, contra Pettit, violate minimal rational choice theory and common sense psychology. According to my interpretation, the best explanation for the plausibility of this goal-displacement has its ultimate grounding in the fourth element of the self introduced above. The prospective father does not literally 'take-in' his child's future goals in the sense of the latter's intentional perspective and stance suddenly replacing his own.[5] Rather, goal-displacement takes place in so far as the father no longer takes his current goal of being an artist as choice-guiding and replaces it, in the course of employing *his own* capacity of reflective scrutiny and reasoning, with the reasons originating from his child's interests and goals. There is nothing esoteric and counterintuitive about transcending self-goal choice according to this reading of Sen. In so far as we attribute a stability-guaranteeing role to Sen's fourth element of the self (i.e., a stable reflective identity), admitting cases of replacing some of one's goals qua basis of choice with those of others does not result in an implausible account of agency characterized by fluctuating and disintegrating intentionality.

4. Negative Goals and Spectra of Permissibility

We have seen that Sen's criticism of narrow self-interest accounts of rational agency has shifted away attention from the wedge that is driven by commitment between individual welfare and choice to the disassociation of a rational agent's choices from her goals. More recently Sen claims that agents are capable of altering their choices quite radically in the face of other people's goals. In fact, and Pettit criticizes this formulation of the argument, individuals may even replace their own goals with those of others. While I defended Sen's notion of goal-displacing commitment against Pettit's critique I am myself puzzled by some of Sen's ambiguous formulations of goal-displacement. The following is, as far as I can tell, Sen's (2002, 35; second emphases mine) strongest statement of what it means to rationally violate the feature of self-goal choice: "Commitment, [. . .] can alter the person's reasoned *choice* through a recognition of other people's goals *beyond the extent to which other people's goals get incorporated within one's own goals* (thereby violating self-goal choice)."

[5] Hans Bernhard Schmid (2005, 57) points out that the view reconstructed by Pettit would have Sen to understand goal-displacing commitment in a "self-eliminative sense": "In this sense, identification is self-defeating, because the very act of identification *presupposes* the very difference in identity that the agent in question tries to eliminate. On this line, there is no way to go beyond self-goal choice, because no matter how far one goes in making somebody else's goals ones own, it is still invariably one's own goals that one pursues." Schmid's idea of a self that eliminates itself gets close to the main idea of identity-reconstitution in the face of radical choices, defended below.

Pettit (2005, 19) takes this quotation as the starting point for formulating his objection according to which Sen is committed to the "claim that we can be the executors of a goal-system that outruns our own goals". In response to Pettit I have claimed that Sen should be interpreted as talking about reflectively endorsing reasons (originating in goals that are not, at the moment of choice, endorsed by the agent in question) here. This response notwithstanding, another problem with Sen's notion of goal-displacement remains, because there are those passages suggesting that some of our choices do not reflect our goals *at all*. This impression is strengthened by Sen's just-quoted insistence that other-regarding behavior in general is possible because of some instances of commitment going "*beyond* the extent to which other people's goals get incorporated within one's own goals". In this section I take issue with this aspect of Sen's account of how self-goal choice can be transcended by committed rational agents.

Now in so far as one's goals are understood in exclusively positive terms (i.e., as states of affairs that we actively try to bring about), Sen's claim, even in its strongest formulation, plausibly suggests that goal-displacing commitment results in choices that are completely detached from an agent's goals that she otherwise holds at the moment of choice. However, this claim about positive goals does not imply that goal-displacing commitment leads to choices that do not reflect *any* of our goals. An agent's goals in the positive sense are the aims and projects she is actively pursuing. As in the examples discussed in *section 3*, positive goals are things such as the pursuit of artistic careers, professional objectives, the aims associated with specific agential roles, etc. The goal of becoming a world-class pianist is a positive goal because it and the many actions constitutive of it are actively pursued by an agent. As such it can be replaced in its entirety by either the agent's new positive goals or by adopting the content (not the intention!) of other people's goals. In so far as agents are influenced by goal-displacing commitment, and replace positive goals of their own with those of others, their choices do indeed no longer reflect their initial (positive) goals.

This picture leaves out a significant portion of what counts as 'our goals' though. Even choices under the influence of goal-displacing commitment reflect some of our goals and hence do not 'outrun our own goals'. These goals are negative ones. Admittedly, the notion of a 'negative goal' might appear paradoxical at first sight. Such goals are not pursuits, endeavors, or projects understood as actively pursued states of affairs. Rather an agent's negative goals are playing the role of constraints, i.e., they set limits on what kinds of positive goals the agent permits herself to get incorporated into the basis of her choice-behavior. Moreover, negative goals are the boundaries that delimit which modifications and displacements of positive goals we permit and which ones we deem impermissible and reject as lying outside of our normative self-conceptions and practical identities.[6]

Returning to the example of the artist helps working out the role that negative goals play in refining Sen's account of commitment. Negative goals define a

[6] The notion of "practical identity" is influenced by Christine Korsgaard's 1996.

certain spectrum of permissible goals that the artist adopts under the influence of goal-displacing commitment. The artist's negative goals (such as, for example, the one of *not* killing innocents) put a limit on what goals of his child he will allow himself to replace members of his current set of positive goals with. In case the child adopts the goal of supporting an international terrorist organization, the former artist will not understand his commitment as unconditional and will not choose on the basis of the reasons that his newly adopted positive goals generate. Rather the child's positive goals that he permits to determine his choices have to remain within a spectrum of (moral) permissibility. Now the point with respect to Sen's above statements is that the goals that remain within his spectrum provide, insofar as the father adopts one or more of them, the basis for choices that *continue* to reflect the father's goals, namely his negative ones. Hence, these choices do not violate self-goal choice if the latter is understood broadly enough.

Let me introduce another example to highlight how negative goals call into question Sen's claim that cases of goal-displacing commitment can result in choices that do not reflect any of the agent's goals. Let us assume I promise to have birthday dinner with you. I propose to have dinner at my place and promise to prepare whatever dish you desire. In fact it's a feature of my birthday present that I award you this latitude. Now on your birthday you show up at my place with a little kitten in a box asking me to slaughter it and put it onto the barbecue. After all, you say, kitten tenderloins are especially tender. It is in situations like these, and in admittedly much more realistic ones, that our negative goals play a critical role. When promising the birthday dinner I was doing so under the proviso of a background-set of negative goals that I hold (and that in this case thought to be intuitive enough not to mention them explicitly) one of them being to lead a life of *not* brutally slaughtering certain sentient beings in order to barbecue them.

At first sight it might seem as if in cases like the birthday-dinner example we award others a kind of blank check. In other words we transfer something like 'goal-determining authority' to other agents. Once I have made the promise it is up to you to decide what we have for dinner. In Sen's terminology, whatever goal (meal) you choose will be incorporated into my set of goals in the sense of providing specific reasons for action as described in *section 3*. However, we award such a blank check within certain limits. The flip side of this argument, and now we are back with Sen's claim that some choices do not "get incorporated within one's own goals", is that my adopting of your goal *does* reflect some of my goals, namely my negative ones. When you ask me to prepare bland and boring oatmeal I may very well think that this is a rather odd choice and that if I had known how you choose I would have proposed something different in order to celebrate your birthday. However, assuming preparing oatmeal does not conflict with any of my deeply entrenched (we'll come back to what this amounts to below) negative goals, my keeping the promise by making your goal the one that provides the reasons for my actions does reflect my negative goals, circumscribing the spectrum within which your odd choice of having me prepare

oatmeal is located. In other words, as long as your goals remain within the spectrum of moral permissibility (which is defined by my negative goals) then even goal-displacing commitment (in Sen's strongest formulation) does *not* result in choices that are completely detached from *my* own goals. My negative goals set limits with respect to your goals that I accept as the basis of my choices.

5. Negative-Goal-Displacement and Identity

I now consider an objection that might be put forward against my claim according to which even in cases of goal-displacing commitment our negative goals continue to be reflected in our choices. What if, the objection goes, I decide to slaughter the kitten in order to barbecue it for your birthday dinner? What if the father adopts his child's goal of becoming an international terrorist and directs all of his subsequent choices in accordance with it? Put more generally, what if agents violate even negative-self-goal choice, i.e., they choose on the basis of reasons provided by other persons' goals that lie outside their current spectrum of moral permissibility?

Returning to the example of Jim and the Indians is helpful in discussing this objection. The point of Williams' (1973, 98–100) example is to imagine a person, Jim, who is asked to perform an action (shoot one innocent Indian in order to save nineteen others) that he deems impermissible. I imagine Jim saying to himself things like, 'I really cannot see *myself* doing this', 'It wouldn't be *me*, shooting this innocent person', etc. However, and here I depart from Williams' description of the scenario, cutting off one of the innocent Indian's toes in order to save the twenty lives is deemed permissible by Jim. Insofar as he is asked to do that the Indians' goal (viz., to survive by means of somebody cutting of an Indian's toe) becomes his own in the sense defended in the preceding two sections. Adopting this foreign goal (despite the fact that it is not a positive goal Jim is embracing enthusiastically) does reflect some of Jim's goals—among them the negative goal of not killing an innocent person.

In an alternative course of events Jim violates negative-self-goal choice and shoots the one Indian, thereby pursuing a foreign goal that lies outside his spectrum of moral permissibility. Now such radical cases of goal-displacement are, *pace* Pettit, psychologically possible. However, Jim's choice occurs on the basis of goals that transgress his negative goals and this constitutes a normatively relevant phenomenon to which I now turn. Again, a modified version of Williams' (1973, 108–118) "integrity objection" that he presents in connection with the Jim-and-the-Indians example fits well with the current discussion of Sen. What is at stake in situations such as Jim's is nothing less than his identity which is fundamentally constituted by his negative goals.

In the next to last paragraph, I deliberately have Jim say to himself the things that he does in order to emphasize the relationship between an agent's negative goals and her identity. Imagining Jim to be a convinced deontologist, regarding killing an innocent person as absolutely impermissible, the dilemma

he finds himself in does not merely challenge some goals he happens to pursue. Much more than that, the dilemma challenges *him* and his self-understanding. Now if Jim decides to shoot the one Indian (thereby violating one of his negative self-goals), this change of mind is best interpreted as a fundamental reconstitution of Jim's identity caused by a deep disruption of his moral belief system. When Jim deliberates about what to do, he is right in concluding 'Me shooting this one innocent person in order to save the other nineteen? No, that would not be *me*!' This is so because his ultimate choice, violating one of his negative self-goals, turns him into a different person, constituted by a different moral identity. Jim's decision to adopt the utilitarian policy is tantamount to a normative self-reorganization.[7]

A caveat is necessary at this point. My talk about negative-self-goal displacement 'disrupting' an agent's identity might suggest that I defend an overly conservative and static view of moral personality, depicting radical moral development as something bad and to be avoided. My approach might make it sound as if I presume Jim's change of mind to be a change for the worse (thereby suggesting the deontological policy to be superior to the utilitarian). I am not suggesting this. The compatibility of my approach with possible instances of religious conversion etc. underwrites my neutrality with respect to the different substantive moral-philosophical theories currently available. Moral identities might change without suggesting that such identity-shifts towards, for example, being a religious believer are morally objectionable. Jim's inner conflict is better understood as unavoidably involving a straining experience and consisting in a process that undermines, at least temporarily, his personal integrity, where such integrity is necessary for existing as a self-guided being that sees herself as *one* coherent person across time.

This caveat also helps to better formulate my worries about Sen's notion of goal-displacing commitment. Sen underestimates the cognitive and emotional cost (not necessarily in a narrow utilitarian sense of 'displeasure') of radically displacing one's own goals in the negative sense that I have been concerned with in this section: Accommodating the exceptional cases of choices and actions, in which agents transgress even their negative moral goals, requires not merely an unproblematic extension of our traditional conception of economic rationality. These cases are very often not experienced as welcomed instances of joyful and enthusiastic commitment. On the contrary, the process of re-constituting one's self in response to Jim-like choice-situations is in a very real sense a '*self*-sacrifice' and results in a truly '*self*less' choice. Again, the inescapability of evaluative language ('sacrifice', 'conflict') should not mislead us here. Transformations from Saul to Paul are *self*-sacrifices and involve conflicts which are relevantly similar to cases such as the one discussed above of the father unconditionally supporting his son's terrorist ambitions. I am here exclusively focusing on a certain species of inner conflict that is present in both types of cases, i.e., the cases of moving from a blameworthy moral character toward a praiseworthy

[7] See Korsgaard 2009 for an account of action as self-constitution that resembles the one presented in the text.

one, on the one hand, and re-constitutions of agency that go in the exact opposite direction, on the other. I am bracketing the more specific moral-philosophical questions here which are concerned with evaluating whether a particular re-constitution of moral identity is a good or a bad one all things considered. My point is more abstract and general: Even Sen's approach does not pay attention to the specific challenge that all cases of such a rare but possible reconstitution pose—independently of whether this reconstitution is a good thing all things considered or not.

6. Conclusion

This paper is characterized by an oscillation between two philosophical positions. This back-and-forth is due to my partial agreement with Sen and Pettit respectively, while at the same time disagreeing with features of their positions. I started off by confronting Pettit's charge, according to which Sen does contradict minimal rational-choice theory and common sense psychology when he claims that self-interest, in the form of self-goal choice, can be violated by committed agents without this violation thereby rendering them irrational. I defended Sen by arguing that Pettit's critique rests on a too literal reading. We do not become passive subjects to other people's goals by 'taking in their intentions', as if these psychological states were imposed on us under hypnosis or through indoctrination and manipulation, i.e., from forces outside of our deliberative standpoint. Rather, we incorporate other agents' goals in the course of our own practical reasoning. Goal-displacement occurs when we accept other people's goals as sources of decisive reasons for choice; reasons that now trump competing considerations that would have originated from our, dearly held and pursued, goals that we considered non-negotiable before. This claim is metaphysically innocent and does not threaten our stable identities as intending and goal-pursuing agents who are capable of reflectively putting other people's goals in place of our own.

This defense of Sen notwithstanding I then argued that Pettit's critique retains some force. There are passages in Sen in which goal-displacing commitment is presented as resulting in choices that fail to exhibit *any* relationship with one's own goals. My notion of 'negative goals' was supposed to establish the claim that some of our goals are always implicated in our choices, even when we fully displace one of our positive goals with those promoted by another person. I concluded with considering the rare but conceivable case of agents who violate even negative-self-goal choice. I tried to sketch how such choices constitute radical shifts regarding agents' identities as cognitively well-integrated agents who exist as one and the same person across time. Both Sen and Pettit, as well as traditional rational choice theory, work under the problematic assumption that we are dealing with individuals that own an unalterable identity that is once and for all defined. This is the reason for why none of these accounts can accommodate the scenarios of displacing negative, identity-constituting, goals: In

these scenarios we are not merely dealing with one stable agent displacing some of her goals; we are dealing with the case of one person's normative identity turning into a different one, as a result of radical goal-displacement.

Where does all this leave us with respect to the debate about economic rational choice theory and Sen's critique of the *homo economicus* paradigm? Pettit's major worry was that violations of self-goal choice contradict rational choice theory. Sen should, according to Pettit, restrict his claims to the unproblematic scenario of goal-modification since the latter is readily compatible with minimal versions of that theory. My introduction of negative goals does not establish a clear-cut answer with respect to the question of whether to side with Pettit or with Sen. On the one hand, I disagree with Sen's position that a person of stable identity can violate self-goal choice and can adopt goals that are *completely* detached from her current goals. Our negative goals always 'color' the choices we make and the actions we perform. Violating self-goal choice is therefore impossible *relative to* a well-unified and integrated person; *a fortiori*, Pettit's minimal rational choice theory is satisfied at least as long as it applies to stable and well-integrated agents, living and acting under normal circumstances. Such agents never fail to choose on the basis of *their* negative goals and, hence, do not violate self-goal choice.

On the other hand, however, I disagree with Pettit's claim that a person is psychologically *incapable* of adopting foreign goals in the more radical sense discussed in *section 5*. Of course, such radical goal-displacement (in my terms, goal-displacement that violates even negative-self-goal choice) constitutes a major challenge to an agent's integrity and her unity as a person. It is only these radical changes that pose a problem for Pettit's minimal rational choice theory and, yes, to common sense psychology. And indeed, as argued above, in so far as negative goals (as constraints on choice) are constitutive of our personal identities, Jim before the deadly shot is not the same person as Jim after the shot. That human beings are capable of undergoing such radical transformations (transformations that, I concluded, are tantamount to reconstituting their identities) is an issue that outruns rational choice and economic theory. And even Sen's broadened conception of rationality has not yet fully appreciated the severity of these exceptional cases that deeply challenge the choosing agent and her*self*.

References

Becker, G. S. (1976), *The Economic Approach to Human Behavior*, Chicago: Chicago University Press.

— (1996), *Accounting for Tastes*, Cambridge/MA: Harvard University Press.

Korsgaard, C. M. (1996), *The Sources of Normativity*, New York: Cambridge University Press.

— (2009), *Self-Constitution: Agency, Identity, and Integrity*, New York: Oxford University Press.

Pettit, P. (2005), "Construing Sen on Commitment", *Economics and Philosophy* 21(1), 15–32.

Schmid, H. B. (2005), "Beyond Self-Goal Choice: Amartya Sen's Analysis of the Structure of Commitment and the Role of Shared Desires", *Economics and Philosophy* 21(1), 51–63.

Sen, A. (1977), "Rational Fools", *Philosophy and Public Affairs* 6(4), 317–344.

— (1982), *Choice, Welfare and Measurement*, Cambridge/MA: MIT Press.

— (1985), "Goals, Commitment, and Identity", *Journal of Law, Economics, & Organization* 1(2), 341–355.

— (1999), *Reason Before Identity—The Romanes Lecture for 1998*, New York: Oxford University Press.

— (2002), "Introduction: Rationality and Freedom", in: *Rationality and Freedom*, Cambridge/MA: Harvard University Press, 3–64.

— (2004), "Social Identity", *Revue de Philosophie Économique* 9(1), 7–27.

— (2005), "Why Exactly is Commitment Important for Rationality?", *Economics and Philosophy* 21(1), 5–13.

Williams, B. (1973), "A Critique of Utilitarianism", in: Smart, J. J. and B. Williams, *Utilitarianism: for and against*, New York: Cambridge University Press, 75–150.