Rational Choice Theory, Commitment, and Toward a Rational Basis for Quitting Smoking

Introduction

Over the past several decades Amartya Sen has argued that the standard behavioral assumptions of economists and decision theorists are impoverished by not allowing for the possibility of committed action. In this chapter, I review Sen's analysis of committed action as it has developed over the years and its implications for rational choice theory (RCT). According to Sen, there is a distinctive attitude that he calls *commitment*, not the same as *sympathy*, which casts doubt on the project of representing economic behavior as the maximization of an agent's utility.

In part I, I trace Sen's analysis of committed action as it has developed over the past years. In part II, I raise objections and consider some recent modifications to his mature position. I argue that Sen's focus on committed action is a curious reemergence of this term, and that it reflects an unconscious, but undeniable commitment to some of the important elements found in Brandom's model of practical reasoning. I conclude that his attacks on RCT can be fought off by recognizing that practical commitments are not competitors to desires and other conative states of mind; instead, commitments are the normative atoms in a rational agent's practical deliberations that desires and other pro-attitudes make explicit. In the final part of the paper, I turn to the Becker-Murphy model of rational addiction. Drawing upon the modifications offered in part II, I argue that addictive behavior is not rational on a simplified Becker-Murphy model, contrary to the authors' own conclusions, because intrapersonal-prudential commitments serve as normative constraints on how much an agent ought to discount future welfare.

Intentional explanation and rational choice theory

Rational choice theory is a variety of intentional explanation.[1] Intentional explanation employs items containing intensional-with-an-s contents such as beliefs and desires (as well as other pro-attitudes, which will always be assumed when I speak of desires). One mark of intensional-with-an-s contents is failure of substitution for co-referring terms. Although Lois Lane loves Superman, she does not love Clark Kent *under that description*, even though both proper names refer to the same fictional superhero. Some theorists have taken this to be one reason for thinking that intentional explanation can never be reduced to the neurosciences. These reference failures, it is said, are a manifestation of the alleged irreducibility of mind to body showing up in the practical realm. Whether intentional explanation will ever be subsumed by the neurosciences or not, we are stuck with it for the time being. Over the last seventy years, since the advent of mathematically formal economics, RCT has been become increasingly refined with the hope of improving its predictive accuracy. In this section and below, I review some of the different senses of *preference* and *utility* commonly used by economists in order to arrive at sharper picture of RCT.

The core of intentional explanation is represented by formula [L]:

> [L]: If any agent, *x*, wants *d*, and *x* believes that *a* is a means to attain *d* under the circumstances, then *x* does *a*. (34)[2]

[L] by itself is inadequate both as an explanation and as a justification of action. At the

[1] My presentation most closely follows presentations of RCT given by: (Elster, 1986) and (Rosenber, 2008)
[2] Rosenberg. Page

very least, we need to add the following conditions:

1. *x* wants *d*.
2. *x* believes that doing *a* is a means to bring about *d* under the circumstances.
3. There is no action believed by *x* to be a way of bringing about *d* that under the circumstances is more preferred by *x*.
4. *x* has no wants that override *d*.
5. *x* knows how to do *a*.
6. *x* is able to do *a*.[3]

With these conditions in place, [L] allows us to infer the missing item in the triad of belief, desire, and action that rationalizes a piece of behavior, given knowledge of two of the items. We can infer that an agent who wants *d,* and believes that *a-ing* is a good way to satisfy his want, will *a*; we can infer that an agent who wants *d,* and is expected *to-a* (however we come to expect this), believes that *a-ing* is a good means to satisfy *d*; and we can infer that an agent who is expected *to-a,* and believes that *a-ing* is a good means to satisfy *d,* desires *d*.

But other conditions on the types of background beliefs and desires that *x* must have can easily be found: *x* believes that *a-ing* is not likely to kill him, *x* believes that he is not restricted from consuming *d* by his probation officer, etc.. In turn, these additional background beliefs imply other background desires: *x* does not want to die, *x* does not want to return to prison. Although the context typically allows us to infer which background beliefs and desires we may assume, this rich background structure assumed with applications of [L] is another statement of the holism of the mental discussed in Chapter 2. Although we may not identify the specific cause of an action with a single belief-desire pair (because we must assume a rich structure of background beliefs and desires), Blackburn explains that we can read off the entire matrix of one's beliefs and

[3] Rosenberg.

desires revealed in behavior "*en bloc.*" This is not so daunting once we realize that the overwhelming majority of our background beliefs and desires are similar in most agents.

Further conditions on intentional explanation can be found within specific theories of practical reasoning. For example, Davidson argues that for something to serve as a reason for action (understood as a belief-desire pair) it must be possible to see in rough outline how that reason caused the action *in the right way*—that is, *qua reason*.[4] This last caveat is needed to rule out cases in which a reason (e.g., a desire to be free of the weight of the climber below) accidently causes an action (the realization that I even have this desire startles me, which causes me to loosen my grip on the rope). As such, we may add:

7. *x*'s belief that *a* is the best means to attain *d* in conjunction with his desire for *d* caused *a*.
8. *x*'s belief that *a* is the best means to attain *d* in conjunction with his desire for *d* caused *a qua* reasons.

Finally, we have conditions emphasized by economists:

9. internal consistency of preferences (a complete and transitive preference ordering).
10. agents maximize their utility.[5]

Condition 10 may be equivalent to condition 4 depending on how we understand "utility." Condition 9 may take different forms depending on how we understand "preference." I give a rough taxonomy of the different senses of preference used by

[4] Davidson.
[5] Sen…*Ethics and Economics*.

economists below, but for our purposes here we can understand a consistent preference-profile to imply the following three conditions:

11. a complete and transitive ranking of one's desires.
12. consistent beliefs.
13. justified beliefs.

Condition 13 in turn comprises the following four claims:

14. beliefs have a maximal degree of inductive plausibility, given the evidence.
15. beliefs are caused by the available evidence.
16. evidence causes the belief 'in the right way'.
17. deductively inferred beliefs may be given as the conclusions of sound arguments.[6]

No doubt, the list is not exhaustive of all possible conditions we might add, but it demonstrates that RCT may be given a very robust interpretation. It includes not only the usual consistency requirements on beliefs and desires, but also epistemological conditions on the types of beliefs that one should have, as well as on the methods by which beliefs ought to be acquired. However, partly because mathematically formal economics first began to develop in the grip of logical positivism, and its offspring—i.e., behaviorism and revealed preference theory—and party because economists (for whatever reasons) are interested almost exclusively in the formal properties of their models, and not the behavioral assumptions that underlie them, a minimal version of RCT has emerged.

The basic idea behind the minimal version is that, with certain conditions met (typically, conditions 9 and 10), behavior can be represented as the maximization of a

[6] Elster.

maximand.[7] The maximand is generically understood as an agent's utility. Unfortunately, the fact that "utility" has acquired different senses as it has come down over the years multiplies the different species of theories all of which fall under the genus RCT.

Broome identifies at least four senses of "utility" in current and past usage going all the way back to Bentham.[8] For Bentham, "utility" means: "that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered." Utility, in this sense, refers to properties of *objects* that tend to produce some good or benefit in a person. A century later, however, we find Edgeworth using the term in a shifted sense (at times) to refer to an agent's *good* or *benefit* itself, e.g., pleasure. A third sense is the one most commonly used by economists and decision theorists today that means an *ordinal representation of preferences*. And finally, there is a fourth sense that is used to represent an abstract representation of an agent's welfare, but in a sense less specific than Edgeworth's or Bentham's understanding of welfare.[9] The third and fourth senses are often used interchangeably because the received view of consumer-welfare in economics is a preference-satisfaction theory. It is the third and fourth senses of utility that will interest us below. But because the third sense employs the notion of a preference, we now have to survey the different senses of preference that economists have on tap.

[7] I will consider these 'conditions' as they come up in the natural flow of the discussion.
[8] Broome, John. In *Ethics Out Of Economics*. 1999. (Cam bridge University Press: Cambridge, UK).
[9] Sen

On the concept of *preference*

In one of Sen's earlier papers, "Behavior and the Concept of Preference," (1973) [10] he identifies three explanatory, and three normative *uses* of preference, which he says can be found to be used interchangeably in the literature. The explanatory uses of preference he identifies are: 1) to describe a person's choices, 2) to represent whatever motives underlie a person's choices, and 3) to represent a person's welfare. He cautions that minor confusions may arise from conflating these uses. I would point out that the third use of preference is identical with the fourth sense of utility mentioned above. I am not sure if any harm can come of this, but it does show how untidy the economist's conceptual tool shed is.

The real danger, Sen warns, comes from the temptation to slide from a descriptive use to one of the following three normative uses: 4) individual welfare as the satisfaction of preferences, 5) overall good of society as aggregate preference-satisfaction, and 6) to articulate a principle of rational choice. It is the slide from 1 to 4 that most troubles Sen. He rightly states that "one is not entitled to infer that a particular choice advanced the individual's welfare just because she made it voluntarily."[11] This injunction is nothing new; it is the one made famous by Hume against inferring ought-statements from is-statements, but cast in the language of preference.

Following Sen's lead, Hausman identifies common *senses* of preference that partly overlaps Sen's list of uses, but also adds to it. Hausman argues that the professional community should embrace a single sense of preference and drop the others in order to avoid inevitable mistakes and confusions. Sen, on the other hand, counsels that we need

[10] Sen, Amartya. "Behavior and the Concept of Preference."
[11] Sen.

to be aware of the different senses, but does not recommend a regimented use of the term because the different senses give economists more tools.[12] In my view, unless we are running out of words, I think that it would be prudent to agree on a single sense of preference.

1. *Choice ranking*.

This sense of preference is the one used by Samuelson in his revealed preference theory. Equating preference with choice allows economists to bypass talk of intentional contents altogether. It is the sense of preference that economists flocked to after Robbins' allegedly devastating critique of interpersonal comparisons of utility.

One objection to it is that some preferences may never reveal themselves in behavior, even over the course of an entire lifetime. I prefer peace over war in the twenty third century, but how could this preference possibly reveal itself in behavior (as opposed to revealing itself verbally) ?[13] This has led some theorists to switch from talk of actual choice to *hypothetical choice*. My preference for peace in the twenty third century could (would?) reveal itself in behavior if certain hypothetical conditions were met—being alive in the twenty third century is a good start. But hypothetical choice runs into problems in game-theoretic settings. In a PD, actors have preferences over *outcomes* ("comprehensive outcomes," says Sen). An actor in a PD prefers the outcome where both play Dove to the outcome where he plays Dove and his opponent plays Hawk. Yet, preference used in this sense is not the same as saying that a player would hypothetically choose Dove if his opponent chose Dove. Hawk is the strictly dominant strategy and

[12] Hausman, Daniel. ….
[13] Hausman's example.

therefore a player should always prefer to play Hawk. Players chose strategies, and prefer outcomes.

2. *Expected-advantage ranking.*

In "Rational Fools," Sen identifies what he calls "the usual sense" of preference: a person prefers *x* to *y* if and only if the person believes he or she will be better off with *x* than with *y*.[14] Whereas the choice interpretation of preference is purely descriptive, this sense of preference is robustly normative. The problem with this sense is that it does not allow us to model agents as satisfying preferences who act in ways that are opposed to what they believe to be in their self-interest, e.g., moral causes. We will return to this conception of preference below, as it does much of the heavy lifting in Sen's early arguments for why committed action presents a problem for RCT. In addition, Hausman comments that what Sen calls "the usual sense" of preference is not one commonly used by economists, but is rather an idiosyncratic use instead (footnote).

3. All-things-considered ranking

Preference understood as an *all-things considered ranking* is the sense that Hausman recommends to the community at large. What to rank can be filled in differently. If it is desires, then a preference is an all-things-considered-desire. If it is reasons—understood differently than as belief-desire pairs—then a preference is an all-things-considered reason. I agree with Hausman that this sense of preference has much to recommend itself. It incorporates the idea of a ranking, which is implicit in the idea of a preference (condition 4). If I prefer a Bottingens, then it follows that I desire it over any other item

[14] I thank Hausman for this understanding of 'preference'.

in my choice set (of equal cost, convenience, at that particular time of day, etc.). It also incorporates the necessary background beliefs that enable [L] to rationalize my choice, e.g., I believe that I will receive a Bottingens if I ask for one, I believe that its consumption does not contribute to the destruction of the environment, etc.. Furthermore, it allows preferences to range over all different items of choice. As such, RCT that embodies this sense of preference is a minimalist version because there is no conceptual connection between the satisfaction of preferences and welfare. I adopt this sense of preference for my analysis below.

RCT: descriptive, predictive, explanatory and/or normative?

RCT may be given either a descriptive or a normative interpretation. It should be clear that one's choice of a particular sense of preference will heavily influence his interpretation. Actually, the order of implication has traditionally run the other way around. Samuelson, and later Friedman, thought that the goal of economic science to be predictive accuracy. For them, economic models are to be accepted or rejected as they are confirmed or falsified over repeated trials. Economics, then, was seen as a purely descriptive enterprise that did not need to *explain-in-the-causal-sense* why agents act as they do. As for normative economics, it was safely quarantined away, and left to work on by adjuncts and other disgruntled employees.

From a purely philosophical point of view, a descriptive interpretation of RCT presents an easy target. This is true partly because of the problems associated with a choice-sense of preference, and partly because of the demise of logical positivism (and its offspring), which has left a descriptive interpretation without a firm foundation on which

to place its behavioral assumptions. In what follows, I follow Sen and others in giving RCT a normative interpretation., although my interpretation of RCT is only weakly normative.

Having said this, I hasten to point out that our interpretation of RCT will likely have only a limited, or even nonexistent impact on much of the vast economic literature. The reason for this is that the two interpretations make different assumptions about the rationality of economic actors. Descriptive theories must assume perfect rationality, where by this I mean behavior consistent with various weaker or stronger axioms. Actors that violate these axioms are not eligible for interpretation. Normative interpretations of RCT evaluate the rationality of actors in light of their behavior. A normative interpretation aims at evaluating the rationality of action.

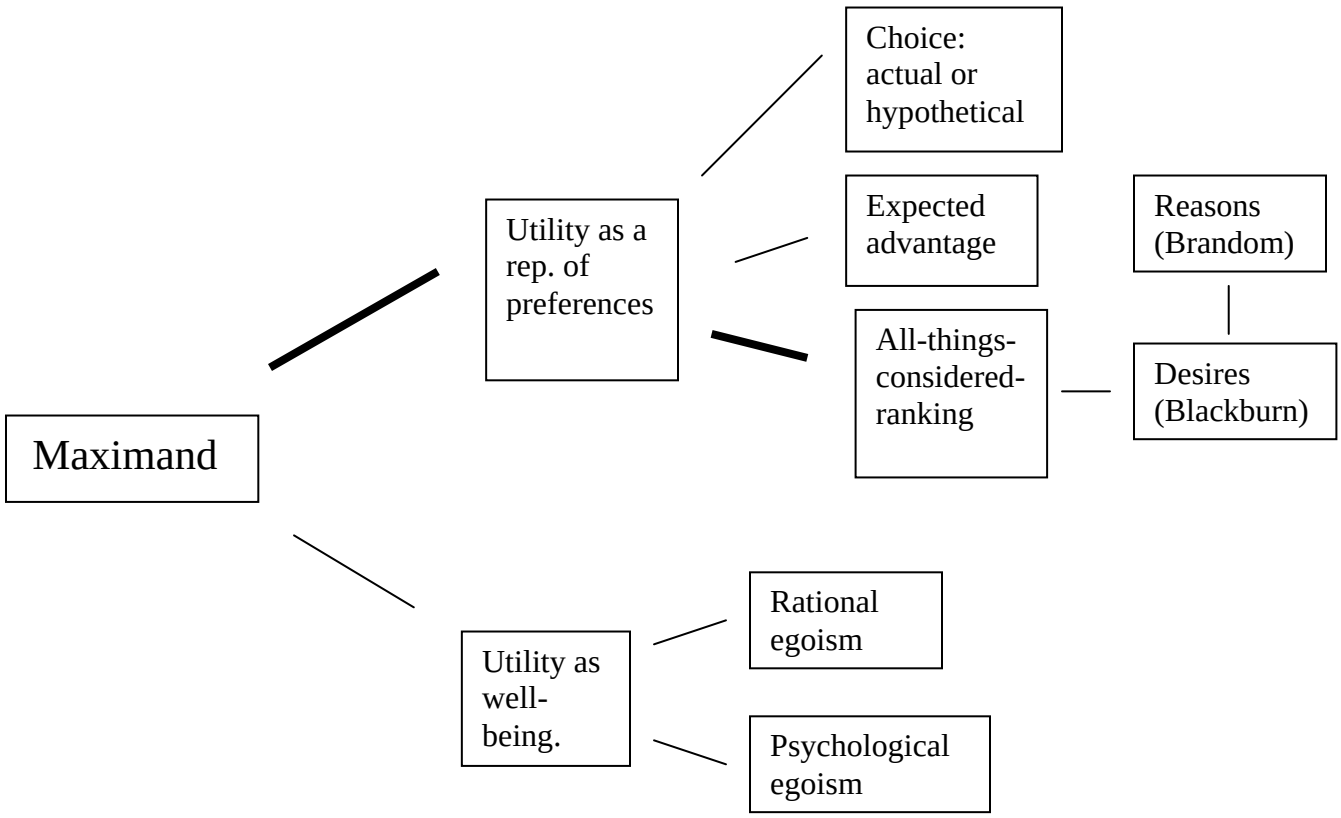
But it may be perfectly reasonable to assume perfect rationality in a wide range of economic settings. For example, when dealing with macroeconomic phenomena we might reasonably assume that individual departures from perfect rationality cancel each other out—that the aggregate quantity is close to what it would have been if everyone had acted perfectly rationally. In addition, the demands of tractability may force us to assume perfect rationality in a great number of settings. Sen comments that these demands present "a hard choice between simplicity and relevance (206-7)."

> We want a canonical form that is uncomplicated enough to be easily usable in theoretical and empirical analysis. But we also want an assumption structure that is not fundamentally at odds with the real world, nor one that makes simplicity take the form of naivety.

Economists should strive for a type of *reflective equilibrium* that aims at balancing the realism of assumptions with the mathematical complexities of model building. One may

reasonably argue that the assumption of perfect rationality is commonly a good one to make because of the demands of tractability. I argue below that at least in the context of addiction, this assumption is not warranted.

To summarize some of the results that will be of use below, I offer the following pictorial overview of the landscape:



The line in bold identifies the path through the trees that I endorse. The basic model of practical reasoning is Brandom's (further articulated with the aid of Bratman's idea of a plan); preference refers to an all-things-considered-ranking; the items that we rank are desires; and utility is an ordinal representation of preferences. In this way, my interpretation of RCT is weakly, but doubly normative. It is weakly normative in that

there is no conceptual connection between preference-satisfaction and welfare. But it is a normative interpretation because it signals departures from perfect rationality both when agents fail to act in light of their best reasons, and when agents fail to appropriately rank their reasons as they ought to do.

Sen on committed action

By "committed action" I will mean action in light of one's commitments. This uninformative preliminary-definition will be filled in as the discussion unfolds. Sen thinks of paradigm cases of committed action as action according to *rules of conduct.* In "Rational Fools," Sen argues that while sympathy may broaden the types of considerations that enter into an agent's practical deliberations by making him sensitive to the welfare needs of others, commitment transcends considerations of welfare altogether. He understands committed action "in terms of a person choosing an act that he believes will yield a lower level of personal welfare to him than an alternative that is also available to him. (footnote)" He allows that there may be cases where committed action coincides with anticipated welfare-maximization. In a limiting case, we can imagine a person whose welfare-maximizing choices are the same as those he would choose in light of his commitments (e.g., a loving saint). To accommodate this (remote) possibility, Sen builds in the caveat that committed action ceases to be welfare-maximizing under at least one counterfactual condition (327). In other words, Sen thinks of commitment as a distinct source of motivation that may compete with other motives in a person's deliberative purview, such as a motive to maximize one's welfare (if there is such a motive).

In this early paper, Sen is using the *expected-advantage-ranking* sense of preference. Given his definition of committed action, it is analytic that committed action drives a wedge between choice and welfare (under at least one counterfactual condition). If committed action is possible at all, and committed action is not welfare-maximizing behavior (under at least one counterfactual condition), then it follows that agents who act on the motive of commitment are making choices that are not welfare-maximizing (under at least one counterfactual condition). His conclusion is that RCT has an impoverished view of human motivation. He cites studies on worker-motivation in Japan, and how standard economic models fall short because they insufficiently account for the loyalty Japanese workers feel toward the firm, where firm-loyalty can be thought of as a type of commitment.

Sen's analysis is fine, as far as it goes. But it does not go very far because of his use of the *expected-advantage-ranking* sense of preference. This turns RCT into a robustly normative theory that counts as irrational any choice that is not welfare-maximizing. In particular, Sen's analysis leaves a minimalist interpretation of RCT untouched, such as one that embodies the *all-things-considered-ranking* sense of preference.

Commitments and Goals

In the 1980's, Sen's account of committed action acquires more sophistication. His main idea is that committed action undermines the commonly-held assumption that agents only pursue their narrow self-interest. One reason that economists assume self-interest is because in basic textbook-explanations of the theory of general equilibrium, market failures—where competitive equilibria are not Pareto efficient—can occur if consumers have interdependent utility functions. Because sympathetic motivation can be

modeled with interdependent utility functions, the self-interest assumption is intended to exclude such disruptive attitudes as sympathy.

Sen explains that the concept of *self-interest* can be broken down into three "essentially independent" features that have traditionally have been lumped together under the same name.[15]

> *Self-centered welfare*: A person's welfare depends only on his or her own consumption (and in particular, it does not involve any sympathy or antipathy toward others).
>
> *Self-welfare goal*: A person's only goal is to maximize his or her own welfare, or—given uncertainty—the expected value of that welfare (and in particular, it does not involve directly attaching importance to the welfare of others).
>
> *Self-goal choice*: Each act of choice of a person is guided immediately by the pursuit of one's own goal (and in particular, it is not restrained by the recognition of other people's pursuit of their good). (footnote)

Departures from self-interest can take the form of violations of any combination of these three assumptions (81). A person violates self-centered welfare by being responsive to the welfare needs of others. A person violates self-welfare goal by failing to act in ways that maximize his welfare. And a person violates self-goal choice by pursuing goals that are not *his own*.

While violations of self-centered welfare have been widely discussed (where?), violations of self-welfare goal and self-goal choice have largely flown under the radar.[16] And while committed action can violate both self-welfare goal and self-goal choice, Sen's focus is on violations of self-goal choice. His claim is that committed action violates this assumption because commitments do not aim at the accomplishment of goals

[15] Sen, Amartya. *Ethics and Economics*. (1988).
[16] Gary Becker. Discussion of violations of the first two assumptions include….. P. 219 of *Rationality and Freedom*

that *belong* to any single individual. The idea rings strange, and I think it is wrong. But hopefully the ensuing discussion will at least make it plausible.

Pettit identifies two senses of commitment that are implicit in Sen's writing. The first sense is what Pettit calls *goal-modifying* commitment. Goal-modifying commitment occurs when someone recognizes that his originally intended goal will adversely affect the welfare of others, and modifies his goal in light of this commitment (…not to adversely affect the welfare of others).

Goal-modifying commitment *does* undermine self-welfare goal because it steers agents away from their welfare-maximizing choice. And goal-modifying commitment also undermines self-goal choice on a robustly normative interpretation of RCT. If we think of goals as success conditions for preferences, and preferences are understood in the expected-advantage-ranking sense, then a modification in one's goals, in one's preferences, should be expected to decrease one's expected welfare, if compelled by an acknowledgement of a commitment (a rule of conduct). But goal-modifying commitment does not undermine self-goal choice on a minimalist interpretation of RCT because the achievement of goals—the satisfaction of preferences—is conceptually unconnected with considerations of welfare. Within a minimalist interpretation of RCT, maximizing welfare with respect to one's goal-modifying commitments is structurally similar to a consumer maximizing utility with respect to his budget constraint. It would be alarming indeed if the latter were not capturable on a minimalist interpretation of RCT.

The second type of commitment Pettit calls *goal-displacing* commitment. These types of commitments occur when an agent not only abandons a goal in light of a commitment—e.g., that his originally intended goal cannot be achieved or even modified

without adversely affecting the welfare of others—but also takes his guidance from that very same commitment, and not from another goal of his own. These types of commitments certainly do undermine the self-goal-choice assumption even on a minimalist interpretation of RCT.

One problem with goal-displacing commitment is that examples are not so easily found that cannot be understood as cases of goal-modifying commitment instead. "I now remember that I promised to keep lookout, so I have to abandon my original goal of practicing yoga today." Can he not just modify his original goal by practicing yoga tomorrow? Another problem is finding examples of goal-displacing commitment where the recognition of a commitment forces an agent not just to abandon his original goal, but also to take direction from that very same commitment, and not from another goal of his own. "I was going to a one-time event, but these people are dying and need my help." Isn't helping these people now his goal? And doesn't this goal *belong* to him? Pettit remarks that action that is not controlled by a goal of an agent is like "trying to imagine the grin on the Cheshire cat in the absence of the cat itself.(21)" It is mysterious how one might intentionally act without a view to realizing a goal.

Commitment and identity

Sen is aware that goal-displacing commitment may strike some as "ununderstandable."(f) He responds by positing a psychological mechanism through which the goals of others may come to be treated *as-if* they were goals of one's own. He calls this mechanism *identification*. Whereas *sympathy* is the mechanism through which the welfare of others comes to be treated as if it were one's own, *identity* is the mechanism through which the goals of others come to be treated *as-if* they were one's

own. Identification can be thought of as a component of a mature agent's conception of his welfare. Where sympathy marks one type of departure from self-interest, commitment marks another (25). And the only limit to whom we may identify with is practical: "Community, nationality, class, race, sex, union membership, the fellowship of oligopolists, revolutionary solidarity, and so on, all provide identities that can be, depending on the context, crucial to our view of ourselves, and thus to the way we view our welfare, goals, or behavioral obligations. (215)"

The addition of identification to an agent's deliberative makeup is included to make plausible the idea that an agent may act intentionally in light of none of his goals: he may be seen as pursuing as-if goals with those whom they identify. As an example, consider a particularly *un*sympathetic individual who drops what he is doing to help those in dire need. He can be seen as displacing his previous goal (whatever it was), and acting now on a commitment (e.g., a basic duty to aid those in dire need). He does not act on a goal of his own, but acts as-if he were. Because he identifies with those in need, he may commit himself to goals that lie outside of his own personal goal-set.

Pettit speculates whether Sen's analysis of committed action reflects a commitment to what he calls the *integrated-deliberation thesis*.

> *The integrated-deliberation thesis*. The schema for non-selfish deliberation, but only deliberation from a limited basis: that of goals that the agent has internalized and integrated into a standing structure. Thus, when one operates in accord with the schema one can only be deliberating on that limited, integrated basis; when one deliberates otherwise—if this ever happens—one cannot be acting in accord with the schema. (ft)

The integrated-deliberation thesis is in keeping with a minimalist interpretation of RCT in that it does not presuppose any conceptual connection between the achievement of goals and an increase of welfare. But it also demands that the goals that *belong* to an agent

must be those that he has "internalized and integrated into a standing structure." This suggests, says Pettit, that goal-initiation is a "distinct psychological episode, one perhaps with a phenomenology of its own." (ft) Is Sen committed to the integrated-deliberation thesis? I return to this question below after I translate Sen's theory of well-being into the language of preference-satisfaction.

Well-being, preference, commitment, and agency

Sen's theory of individual well-being starts with idea of a *functioning achievement*. A functioning achievement can be understood as a *being* or a *doing* that one has reason to value: being well-nourished, going to work, feeling content, practicing the oboe, etc.. Although functioning achievements constitute part of a person's *well-being*, Sen reserves the term as a measure of what he calls a person's *capability set*. A capability set is an agent's set of functioning achievements. Because functioning achievements are things that a person may do, and the ways he can be, an increase in a person's capability set adds to the number of things he can do, and the ways he may be. And because Sen defines functioning-achievements as *doings* and *beings* that an agent has *reason to value*, adding functional-achievements to a person's capability set increases the number of *valuable* things he can do, and the *valuable* ways he may be. In this way, Sen holds that a person's capability set marks out a special sense of *freedom*, or what comes to the same thing, is a measure of a person's *well-being*. The person who is fasting is much better off than the person who is starving, even though both are experiencing the same level of biological functioning (malnutrition)—Sen assumes this to be irrefutable. And the reason that fuels this intuition, concludes Sen, is that the person who is fasting is free-in-this-special-sense to eat, whereas the person who is starving is not.

Sen's theory of well-being can be translated into the language of preference-satisfaction with the aid of the idea of a *content-independent-decisive preference* (CID-preference). A preference is said to be decisive if it is *possible* for it to be the reason-*in-a-causal-sense* for why an agent acts. If I prefer A over B, then my preference is decisive if it is possible for it to be the reason why I choose A over B, even if I never choose A over B. If I am cold and prefer that the door be shut, then my preference is decisive if it is possible for me to shut the door *because* I prefer it to be closed. It is not necessary for a preference to be decisive for it to have been a reason for action in the past. Sen calls this a case of decisive-choice. It is necessary only that it *may possibly* serve as a reason for action.

A preference is content-independent-decisive if its content does not determine whether it is decisive or not. If I prefer A over B, then my preference is content-independent-decisive if my preference may serve as a reason for choosing A over B (even if I never do), provided that I am free to choose B instead (its content does not rule out the possibility of it being chosen). My preference for coffee over tea is content-independent-decisive if it is possible for it to be the reason why I chose coffee over tea, provided that tea is available as well. A prisoner's preference to remain locked up over going free is not content-independent-decisive because freedom is not an option for him.

Functioning-achievements can be thought of as content-independent-decisive preferences, with the important caveat that we are using preference in the expected-advantage-ranking sense. The reason for this is because functioning-achievements (by definition) refer only to things that we have reason to value. Thus, of the three senses of preference discussed above (choice-ranking, expected-advantage-ranking, and all-things-

considered-ranking), only expected-advantage-ranking fits this bill, because only an expected-advantage-ranking sense of preference contains a conceptual connection between the satisfaction of preferences and welfare. The conclusion: Sen's theory of well-being can be understood as a measure of one's set of content-independent-decisive preferences in the expected-advantage-ranking sense of preference.

We have left only to see where committed action fits into this framework. Sen thinks of rational agency as comprising both an *agency*-aspect and a *well-being*-aspect. It is unclear to me what exactly Sen means by these terms, but I think that we can understand them as designating distinct sources of motivation. An agent exercises his agency-aspect whenever his motive is commitment. An agent exercises his well-being-aspect whenever he acts on a content-independent-decisive preference. This understanding resonates well with Sen's talk of commitments as motives battling it out with other motives in one's practical deliberations.

Part 2: Objections and Additions

I believe that Sen's analysis of committed action is flawed for reasons that I offer presently. I next consider additions and modifications that are intended to bring Sen's important insights on committed action more in line with these objections and modifications.

The integrated-deliberation thesis

I mentioned above that Pettit thinks that Sen's analysis of committed action expresses a commitment to what he calls the integrated-deliberation thesis.

> *The integrated-deliberation thesis*. The schema allows for non-selfish deliberation, but only deliberation from a limited basis: that of goals that the agent has internalized and integrated into a standing structure. Thus, when one operates in accord with the schema one can only be deliberating on that limited, integrated basis; when one deliberates otherwise—if this ever happens—one cannot be acting in accord with the schema. (ft)

Sen's talk of goals needing to become "incorporated" into a person's standing structure of goals suggests that Pettit is right. Furthermore, without some way of demarcating self-goals from as-if-goals, committed action cannot be seen as undermining self-goal choice. It does appear that Sen is committed to some version of this deliberative scheme.

I think that Sen's commitment to the integrated-deliberation thesis partly stems from a desire to avoid a revealed preference theory. If an agent's goals are thought to be whatever ends he reveals himself to be pursuing in action, then this might seem to come dangerously close to a revealed preference theory. Recall that the sense of preference used in revealed preference theory is choice-ranking. My impression is that Sen believes that the only alternative to choice-ranking is expected-advantage-ranking. Some evidence for this can be gleaned by recalling Sen's cataloguing of the different explanatory and normative uses of preference. One takes away from this list that the only *senses* of preference available to economists are choice-ranking and expected-advantage-ranking. To avoid choice-ranking, Sen embraces expected-advantage-ranking.

If my diagnosis is correct, then Sen understands the *belongingness-condition* on goals to mean that only those goals whose attainment is expected to yield an advantage are goals that can properly be said to be *belong* to an agent, i.e., they are self-goals. Agents who pursue goals whose attainment is not expected to yield an advantage (under at least one counterfactual condition), are not pursuing goals that belong to them. They are

pursuing as-if goals, and they violate the self-goal-choice assumption. Sen's division of rational agency into a well-being-aspect and an agency-aspect reinforces this diagnosis.

But the problem with this understanding of the belongingness-condition is that it does not give us any principled reason for *why* goals whose attainment is not expected to yield an advantage to an agent should not be included in his goal-set. We can just as easily say that goals whose achievement make an agent blush, or giddy, or mad should not be included in his goal-set. On what other grounds might we separate self-goals from other goals that agents pursue?

One way forward is to understand as-if goals as *shared or collective goals*. Much of Sen's discussion of committed action is within the framework of a prisoner's dilemma. He thinks of cooperative behavior in these contexts as committed action that overrides rational actors' strictly dominant strategies. In this way, the cooperative outcome in a PD can be thought of as a collective goal; rational agents in PD's act as-if the cooperative outcome is one of their goals.

Schmid understands Sen's analysis of committed action is this way. He states that "the self-goals which individuals choose when they act together cannot be adequately represented within an account which takes all goals to be self-goals, because these self-goals *presuppose* shared goals. (59-italics mine)" The charge is that shared goals undermine self-goal choice because self-goals (typically, or at least sometimes) presuppose shared goals, and by definition shared goals are not self-goals.

But the claim that one goal *presupposes* another goal can be understood in at least three ways: 1) that it is logically impossible for an agent to have certain goals without also having those goals that it presupposes, 2) that certain goals—those goals that are

presupposed—cause an agent to have other goals, or 3) that certain goals—those goals that are presupposed—imply that an agent *should* (normatively speaking) adopt other goals. By "presuppose," Schick means that self-goals stand in *normative* relations with collective goals. Our shared goals imply that I should adopt and pursue various self-goals. If our shared goal is to get the fuel-less car to the top of the hill, then I ought to have as a self-goal: *push on the car*. If our shared goal is world peace, then I ought to have as a self-goal: *structure my life in ways that contribute to conflict resolution*.

But it is far from obvious why shared goals undermine self-goal choice. For one, it seems entirely plausible to understand one's shared goals also as one's self-goals (for the purposes of RCT, at least). If our goal is to get the car to the top of the hill, then *my goal* is to get the car to the top of the hill. Just because the achievement of a goal is not entirely under an agent's control does not rule it out as a self-goal. The archer's goal is to hit the bull's eye even though this is not entirely under his control. Even my goal of acquiring a turkey sandwich for lunch requires the cooperation of sandwich makers, turkey farmers, etc.. Instead of saying that an agent has self-goals on the one hand, and shared goals on the other, it seems just as reasonable to say that the attainment of many of *his* goals requires the cooperation of others.

Secondly, it unclear of what value knowledge of shared goals is to a rational choice theorist. Certain self-goals *intrapersonally* presuppose other self-goals. Recalling Bratman's discussion in Chapter 3, rational agents have a hierarchy of goals or plans where shorter-term goals presuppose longer-term goals. Must a rational choice theorist have total knowledge of one's intrapersonal hierarchy of goals in order to rationalize his behavior? Surely not. It is enough to know that someone believes that he is Napoleon to

rationalize his peculiar wardrobe choices. It is unnecessary to know whether he came by this belief from a knock to the head, or he infers it (rightly or wrongly) from other beliefs *insofar as those other beliefs are not needed to rationalize his behavior*. Similarly, I fail to see how collective goals are needed to rationalize a person's behavior if we are to think of shared goals as different from self-goals (e.g., it cannot be a self-goal of mine to get the car up the hill). It is sufficient that we know only those self-goals on which an agent acts.

Thirdly, even if shared goals undermine self-goal choice, this might not fit very well with other parts of Sen's theory. If as-if goals are shared goals, then committed action aims at the attainment of these shared goals. But Sen thinks of goal-displacing commitment—as opposed to goal-modifying commitment—as action in the absence of any goal. If we are to say now that goal-displacing committed action aims at shared goals (as opposed to self-goals), then this is a new addition to his theory. The problem with this approach, however, is that shared goals are typically adopted because the participants in a collective venture typically expect to receive an advantage from the attainment of their shared goals. For example, Sen argues that it is rational for rational agents to act cooperatively in a PD *because* these actors are expected to be better off with the cooperative outcome than the outcome that would arise if each followed their strictly dominant strategy (in the absence of consideration of commitment). In other words, joint enterprises, collective goals, are typically justified in terms of welfare. But Sen is very clear that agents who act in light of goal-displacing commitments should *not* expect a welfare return—goal-displacing commitment "transcends" considerations of welfare altogether. They express values for items other than welfare, e.g., environmental values.

Goal-displacing committed action reflects motives associated with a rational agent's agency-aspect, as opposed to his well-being-aspect.

One way forward is for Sen to embrace instrumentalism. In one of the most well-known and influential statements on the methodology of "positive economics," Friedman argues that the *realism* of the behavioral assumptions of economic theory is unimportant to its final aim: predictive success.[17] Behavioral assumptions are good, appropriate, useful, etc., insofar as the theories within which they are embedded prove to be predicatively accurate—more accurate than rival-theories that include different behavioral assumptions, at least. For example, the assumption of self-interest is a good assumption if and only if the predictive success of theories that assume it can be shown to be superior to competitors that include altruistic assumptions about agents. That some people do in fact act on altruistic concerns is irrelevant to theory-choice. Usefulness, and not realism, is the key criterion (but there may be other considerations important for theory-choice as well, e.g., simplicity).[18] In this way, economists, says Friedman, talk *as-if* consumers are, e.g., self-interested; it is a convenient fiction whose usefulness can only be determined through experience and comparison with rival theories.[19]

Should we understand Sen's discussion of as-if preferences as a commitment to instrumentalism? This would have the advantage of making sense of his remarks on the belongingness of goals. We may talk of agents as-if they act, at least occasionally, on goals that do not belong to their standing structure of goals; we can do this without undertaking any serious commitment to statements about how rational agents really are.

[17] "The Methodology of Positive Economics."
[18] It is unclear what Friedman means by "realism." To name just two possible meanings, he may mean ….See Caldwell and other guy
[19] Friedman is a Popperian falsificationist.

But it seems unlikely that Sen would, or should, go along with this instrumentalist approach. After all, Sen's analysis of committed action is meant to draw attention to certain *unrealistic* assumptions doing damage in economic theory right now. In "Rational Fools," he motivates his analysis with a discussion of worker motivation in Japan. The upshot is that the narrow assumptions of standard economic theory are unable to capture these important sources of worker motivation.[20] Furthermore, appreciating commitments as an addition to the limited motivational resources available to *homo economicus* has implications for Sen's unique understanding of agent-welfare, discussed below. As such, instrumentalism would seem a strange bedfellow for Sen.

Commitment and desire

In the introduction, I said that Sen thinks of commitment as a "distinctive attitude," an independent source of motivation not reducible to underlying desires and other pro-attitudes. Blackburn, we saw, understands desires in a causal-functional sense in that our interpretation of an agent must be re-evaluated whenever he fails our expectations. I criticized Blackburn's theory in Chapter 2 for not being able to reduce commitments to desires and other pro-attitudes. Brandom, on the other hand, understands our use of desires in rationalizing behavior as making explicit material proprieties of practical inference—practical commitments. For Brandom, commitments are thoroughly normative entities ("norms all the way down").

The problem I see for Sen is that he appears to be on the fence in terms of the way that he understands the relationship between commitments and desires. If he understands desires along the lines of Blackburn, then commitments are not "distinctive attitudes;"

[20] Worker motivation.

instead, we must find a way of thinking of them in terms of underlying desires. But if we follow Brandom in taking commitments (and entitlements) as the fundamental (normative) atoms in one's practical deliberations, then desires can no longer be seen as "distinctive attitudes," alongside of one's commitments. I am assuming here that one is committed to a model of practical reasoning that is essentially like that of either Blackburn's or Brandom's. I can offer no specific argument for this assumed dichotomy, and so my conclusions will turn on this assumption. But I do think it is odd to view commitments as distinctive attitudes alongside desires, if desires are thought of causal-functionally. And if we are to think of desires normatively, this suggests an approach like Brandom's, especially in light of the fact that Sen is arguing that the possibility of *committed* action undermines self-goal choice—the very notion that figures centrally in Brandom's model of practical reasoning.

If Sen is committed to the claim that rational agents act on commitments at least some of time, and commitments are not to be understood in terms of desires—but instead desires make explicit an agent's practical commitments—then Sen is committed to the claim that all rational action is committed action (where the "is" is predication, and not identification because committed action can of course be irrational). Putting it all together, we can understand committed action as follows: preferences are all-things-considered rankings; what we rank are desires; and desires make explicit practical commitments. As such, committed action presents no more of a problem to self-goal choice than does the idea of agents acting on belief-desire pairs.

Given these remarks, it is still open to Sen to think of violations of self-goal choice as marking out a certain class of commitments. We can regard shared goals as violations

self-goal choice, if we want. However, this would beg the question of why shared goals ought to be excluded. Or, following Brandom's classification of commitments into preferential, prudential, and moral, Sen may think of moral commitments, or a sub-class of our total set of moral commitments, as undermining self-goal choice. This would be consistent with his repeated claims that action in light of commitments is not expected to be welfare increasing, although it may be with a highly sympathetic persons. In the end, I hope that my analysis at least makes clear the need to justify the self-goal-choice component of the self-interest assumption. If excluding shared goals or moral goals is required to avoid market failures, then this would certainly call into question the desirability of market successes.

Toward a rational basis for interpersonal-prudential commitment

In part III, I argue that *intra*personal-prudential commitments can be thought of as constraints on how much a rational agent ought to discount his future welfare. Because I model *intra*personal-prudential commitments on *inter*personal-prudential commitments, and I argue that addiction is irrational on a simplified Becker-Murphy model, I must give some way of understanding the rationality of the undertaking of interpersonal-prudential commitments. By "intrapersonal-prudential commitment" I will mean a commitment whose acknowledgement (or failure to acknowledge) is likely to substantially affect an agent's expected welfare over a period of time, but is not expected to substantially affect the welfare of others (in the complete absence of sympathy). In contrast, acknowledgement of moral and other interpersonal commitments (or failure to acknowledge) can be expected to substantially affect the welfare of others. By "interpersonal-prudential commitment" I will mean the type of commitment agents

undertake in conjunction with other against specifically because they expect a benefit that the attainment of such shared goals will yield. I recognize that some theorists think of moral obligations as comprising only these types of commitments. The distinction is only meant to allow for the possibility of other foundations of moral commitment not based on a contract-model (e.g., as from an exercise of one's pure practical reason).

Recalling my discussion of Sen's understanding of *identity* discussed above, Anderson understands the rationality of what I called interpersonal-prudential commitments in terms of the rationality derived from an agent's practical identities. By "practical identity," Anderson (following Korsgaard) means the different (perhaps, non-moral) ways in which people may identify or associate with one another. Her basic idea is that (certain types of) committed action is based on reasons that it is rational for *us* to adopt—that is, any group of people regarded as a collective agent.[21] This idea is captured by what she calls "*The Priority of Identity to Rationality Principle"*:

> …what principle of choice it is rational to act on depends on a prior determination of personal identity, of who one is.[22]

Consider again the voting-example discussed in Chapter 3. Voting is an n-person PD because the marginal causal impact of a single individual's vote is near zero. Because a voter's contribution to this public good—democratically elected officials—is outweighed by the inconvenience of going to the polls, if a potential voter anticipates almost any inconvenience at all, the principle of expected utility-maximization recommends staying at home. However, the utility-maximizing-reason that recommends staying at home,

[21] Anderson page 24.
[22] Anderson page 30

argues Anderson, is not a reason that we could *accept* other rational agents acting upon. She writes:

> The key to figuring out this [cooperative] outcome [to a PD] is that it is a **constitutive principle** of a collective agent (a 'plural subject' or 'we') that whatever can count as a reason for action for one member of the collective must count as a reason for all. That is, in regarding themselves as members of a single collective agency, the parties are committed to acting only on reasons that are universalizable to their membership. (29, **bold** mine)

Her key move is the claim that it is a "constitutive principle" of practical identification that we regard good reasons-for-action from an individual's point of view, as reasons-for-action that can be accepted from any person-to-the-contract's point of view. As such, the principle of expected utility-maximization in PD-type cases does not yield reasons that are *universalizable*. They are not reasons that we should expect every reasonable person reasonably to accept (or not reasonably reject, or some variant of this idea). What rationality requires here, says Anderson, is a non-act-consequentialist principle of choice. This is not to say that an act-consequentialist principle of choice, e.g., expected utility-maximization, is never the rational principle of choice. On the contrary, for most of our day-to-day decisions, it is just fine, rationally speaking, she says. But when it comes to collective goals, our self-interested reasons may be overridden. And the reason for this is because *what it means* to take oneself as a member of a collective *just is* to recognize certain collective-goal-promoting reasons as having weight. These reasons would not be weighty to those outside of the collective; moral reasons should by weighty to all rational beings.[23] In this way, group-identification is analogous to sympathy: we have reason to promote the welfare of those with whom we sympathize, and we have reason (it is

[23] She borrows from Korsgaard.

rational) to contribute to the achievement of collective goals by acting in committed ways with those whom we identify.

As such, committed action—and in particular, action that is not utility maximizing—derives from the appropriateness of our many socially desirable goals. Of course, an agent need not have such goals consciously in mind. This is only an explanation of the logical order of goals, identities, and commitments.

An immediate problem with this approach is that Hume believes that no rational basis for sympathy can be found. It is generally a good thing for us that others have some measure of sympathy. But what we cannot do, says Blackburn, is "…argue a sensitive knave into upright behavior." We can offer no *rationally* compelling reason or argument for why he ought to be more sympathetic. Sympathy must instilled in someone through a proper moral education, habituation, etc.. Are we to say the same about group-identification? Is it rational to identify with certain groups and not with others? Or is group-identification merely conventional, and without a rational basis? If the rationality of action based on interpersonal-prudential commitments derives from the rationality of our various associations or practical identities, and no rational basis for group-identification can be found, then the search for a rational basis for interpersonal-prudential committed action appears to wash away under our feet.

I do not think that a rational basis for sympathy can be found without engaging central epistemological assumptions that lie at the heart of Hume's empiricism. Furthermore, it is unclear to me whether our understanding of sympathy has any strict implications for what our view of the rationality of our other practical identities should be. I will largely skirt these ominous questions, even though my rejection of Blackburn's Humean model

of practical reasoning in Chapter 2, and my endorsement of Brandom's Kantian-inspired model in Chapter 3, would constitute part of my attempt at an answer. Instead, I offer a start to a sketch for how the rationality of our practical identities may be understood that draws upon Sen's understanding of the role that *discussion* plays in the formation of practical identities.

Borrowing from other themes in Sen's writings, Anderson thinks that *discussion* plays an important role in group-identification. She says that practical identification does not require any prior acquaintance, only that "we see ourselves as solving a problem by joining forces."[24] A "shared intention is sufficient to constitute individuals as a social group with a common practical identity," with the caveat that the only constraint on whom one may share an intention with is that such sharing is possible (31).[25] In this way, through discussion, agents come to recognize certain ends as socially desirable. A consensus on socially desirable goals allows agents to identify with each other. Because of this newly emergent practical identity, it is rational for agents to undertake commitments that aim at the accomplishment of shared goals. In summary:

1. Through discussion comes the recognition of socially desirable goals.
2. We come to identify with others who also find these goals to be socially desirable.
3. We commit ourselves to appropriate shared-goal-promoting courses of action.

On this order of explanation, the rationality of interpersonal-prudential commitments derives from our practical identities—it is rational for us to form practical identities with those whom we share a view to the desirability of various social ends.

Intrapersonal-prudential commitment and personal identity

[24] Anderson p. 31
[25] Anderson p. 31

The problem with Anderson's approach, however, is that it is plausible to change the order of explanation and think of commitments as coming first, prior to the formation of one's associated practical identities. Anderson indeed suggests just this with her statement that identities are *constituted* by commitments. But, if commitments constitute our practical identities, how can a practical identity provide a rational foundation for the undertaking of those very same constitutive-commitments? What at bottom provides a rational basis for our most basic identities? Perhaps commitments are the normative atoms out of which our many identities emerge?

Fortunately, for my purposes below, I do not have to answer exactly these questions, although I venture an answer partly in order to place the necessary analytical tools on the table for the discussion to come. As a preliminary remark, I point out that the undertaking of further commitments—commitments that are not *constitutive* of a practical identity—can be justified in terms of that practical identity. For example, it may be rational to commit oneself to voting for a particular union representative *because* one identifies with union employees and thinks that he is the best candidate for the job. This commitment is not constitutive of his union-identification.

But for constitutive-commitments, I believe that it is plausible to think of practical identities as ultimately emerging from individuals' *personal identities*. I am using "personal identity" here in a specific sense to refer to a rational agent's recognition that he is a *planning creature*—that he ought to set longer-term goals and figure out plans that he can reasonably expect will attain his goals. I noted above that Sen thinks of paradigm cases of commitments as rules of conduct. If we think of rules of conduct as a type of contract (whether implicitly or explicitly, voluntarily or involuntarily, entered into), then

commitments can be understood as the terms of a contact. They specify individual responsibilities in the pursuit of collective goals. But how are we to think of intrapersonal-prudential commitments in the absence of third-parties?

We can extend the contract-model of interpersonal-prudential commitment to cases of intrapersonal-prudential commitment by thinking of the latter variety of commitment as a type of contract that one makes with a future self. The idea is not so mysterious if we recall Bratman's planning theory of agency. Intentions express commitments, and plans ("intentions writ large") are constituted by commitments. The reason we set plans is that they structure our lives. This allows us to pursue longer-term goals—goals that we would unlikely be able to accomplish in the absence of well-structured plans. And the accomplishment of longer-term goals—at least those of a prudential variety, as opposed to moral causes—tend to make us better off *in the long run*.

I think that it is an essential component of rational agency that one identifies with his future selves. Because of this, we commit to goals and plans largely because they are expected to make our future selves better off. But most goals require the cooperation of others. In this way, commitments that constitute our plans will typically include interpersonal commitments as well. As such, the entire network of a rational agent's commitments, and the practical identities that certain commitments constitute, ultimately rests on one's personal identity in the sense I mentioned above. I believe that this view can be augmented by including as part of rational agency (personal identity) the demand that rational agency requires of agents the recognition that the welfare of others is equally valuable as their own welfare. This additional sense of personal identity can be thought of as marking out one's *moral identity*. Commitments justified in terms of one's moral

identity may constrain the other types of commitments that a rational agent should undertake.

In light of these remarks, I disagree with Anderson in holding that the principle of expected utility-maximization should be thought of as a distinct principle of rational choice that is appropriate in certain contexts, but not in others (although I will continue to talk of it as a principle of choice). Because I believe that rational action is committed action, I believe that it is better say that rational agents have *reason* to maximize their expected utility in the absence of certain constraints. These constraints may take the form of interpersonal or intrapersonal, moral or non-moral, commitments. These commitments are justified in terms of an agent's various practical and personal identities.

Even if I have overstepped here, I will now make a case that intrapersonal-prudential commitments can at least serve as constraints on *how little* one ought to value the welfare of his future selves.

Part III: Intrapersonal-prudential commitments and rational addiction

The Becker-Murphy model of rational addiction

Understanding addiction to substances, of course, enjoins many disciplines. I am concerned here with the narrow question of whether addictive behavior can be viewed as rational according to standard RCT. Becker and Murphy argue that addiction can be provided a rational basis.[26] They argue that addictive behavior can be characterized as unstable equilibria at which consumers either increase their consumption over past use—this captures the idea of a developing tolerance toward an addicting substance—or

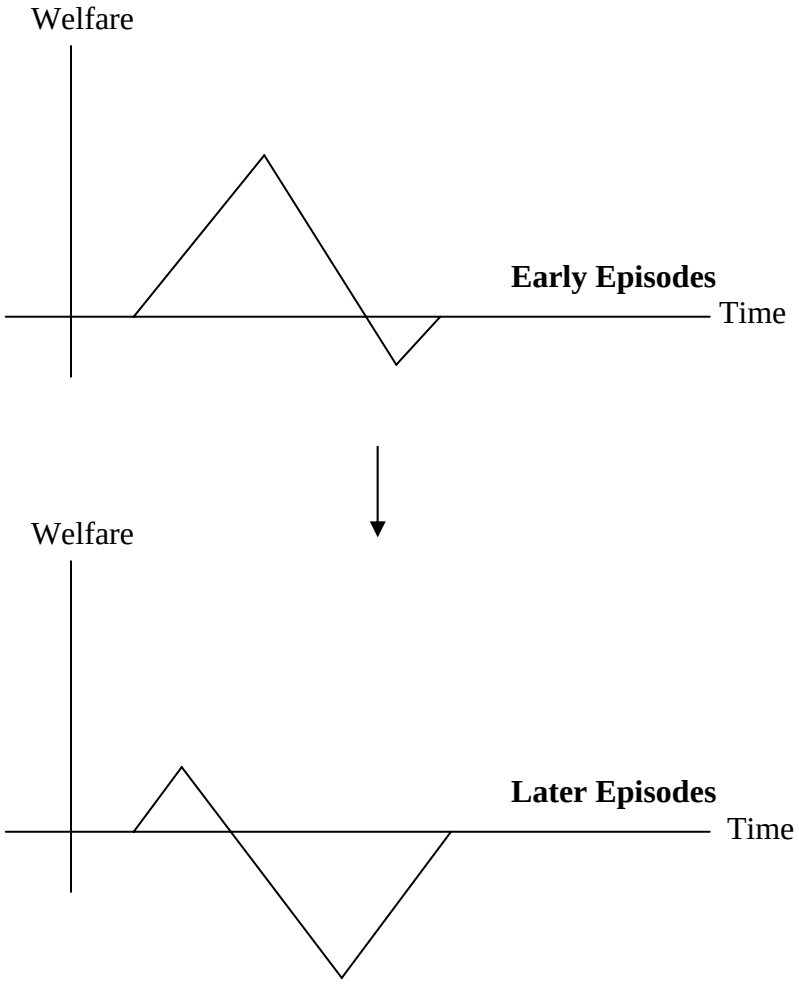
[26] Becker, Gary and Murphy, Kevin …. "A Theory of Rational Addiction" *The Journal of Political Economy*. Vol. 96, No. 4 (Aug. 1988), pp. 675-700

dramatically decrease their consumption as compared with past consumption—this is intended to model quitting "cold turkey." In terms of behavior and utility, an addicts standard progression is one of increasing marginal utility—this represents increasing severity of withdrawal symptoms—but decreasing overall utility—this represents an increasing tolerance of the substance. Because the Becker-Murphy model is largely a complicated mathematical exercise, I will focus on a simplified version of their model provided by Ole-Jørgen Skog. The overarching question in the debate is whether the standards of rationality are the same for time-dependent preferences as they are for time-independent preferences.

Addictive behavior reveals a preference-profile that has the following three properties:

1. *Tolerance*. Consumers need to consume larger amounts of a substance after periods of use in order to achieve the same level of welfare as they achieved with first-time use.
2. *Reinforcement*. After periods of consumption, an agent requires more of the substance to keep up with the joneses.
3. *Discounting*. Consumers discount future welfare (consumption). Although there are compelling arguments against the discounting-assumption, I will not weigh in on those arguments here. This assumption is firmly entrenched in the economic literature, and addiction can hardly be seen as rational without it. But I will say in its behalf that discounting future welfare maps on nicely to economists' notion of the present real value of income (wealth). The present real value of income is deemed greater today than it is at any time in the future because consumers may lend out their present income at a premium. Since income is an important social primary good in the determination of one's welfare, and because the real future value of wealth is a decreasing function of time, it follows that there is reason for consumers to discount their future welfare.

The basic structure of addiction can be represented pictorially in the following two-period model:



In early episodes of consumption, an agent receives a greater benefit from consuming a given amount, and experiences a lesser withdrawal. In later episodes, this same amount of consumption yields a smaller benefit, and an agent experiences a greater withdrawal. The change reflects his increasing tolerance and his more pronounced symptoms of withdrawal (reinforcement). The main idea behind the B-M model is that because future welfare is discounted, it *may* be rational for an agent to continue to consume, or begin to consume even if he has abstained in the previous period. The qualifier "may" is needed because it depends on the degree to which an agent develops tolerance to a substance, the

degree to which his withdrawal symptoms emerge, and the degree to which he discounts future welfare (consumption).

Numerically, we can represent a case of addiction as follows. In keeping with the discussion above, I understand the numbers to be an ordinal representation of a consumer's preferences, and preferences in the all-things-considered-ranking sense. Because I assume that consumers are not ranking their preferences wrongly (as this sense of preference allows), a consumer's utilities can also be thought of as his welfare (in the abstract sense, the fourth sense, discussed above). In this way I am free to switch between talk of utility and welfare. Consider first a case in which an agent does *not* discount future welfare at all.

Time	1	2	3	4	···
Abstain	80	80	80	80	···
Consume	100	70	70	70	···

It is rational for him to abstain in period-1, and continue to abstain in each successive period. His expected future welfare is greater on this path than it is with any other combination of abstention and consumption.

But now consider two different cases: one in which a consumer discounts future welfare at a rate of 60%, and one in which a consumer discounts at a rate of 90%. Discounting is typically represented as a decreasing exponential function of time. For simplification, the discounting-numbers below are approximations of this more complicated formula.

<u>60% discounter starting off as abstainer</u>:

Consumption:
$100 + (0.60 \times 70) + (0.36 \times 70) + (0.216 \times 70) + \cdots = 205.$

Abstention:
$80 + (0.60 \times 80) + (0.36 \times 80) + (0.216 \times 80) + \cdots = 200.$

It is rational for this agent to begin to consume and continue to consume in all future periods, as his expected future welfare is greater on this path.

90% discounter starting off as abstainer:

Consumption:
$100 + (0.90 \times 70) + (0.81 \times 70) + (0.729 \times 70) + \cdots = 730.$

Abstention:
$80 + (0.90 \times 80) + (0.81 \times 80) + (0.727 \times 80) + \cdots = 800.$

Unlike a 60% discounter, a 90% discounter should continue to abstain in all future periods.

Now consider the case when they start off as consumers.

60% discounter starting off as consumer:

Consumption:
$70 + (0.60 \times 70) + (0.36 \times 70) + (0.216 \times 70) + \cdots = 175.$

Abstention:
$40 + (0.60 \times 80) + (0.36 \times 80) + (0.216 \times 80) + \cdots = 160.$

90% discounter starting off as consumer:

Consumption:
$70 + (0.90 \times 70) + (0.81 \times 70) + (0.729 \times 70) + \cdots = 700.$

Abstention:
$40 + (0.90 \times 80) + (0.81 \times 80) + (0.729 \times 80) + \cdots = 760.$

We have the same conclusion. The 60% discounter ought to continue to consume in all

future periods, and the 90% discounter ought to abstain in all future periods.

With a 70% discounter, his decision to consume or abstain depends on whether he consumed or abstained in the past-period.

70% discounter starting off as abstainer:

Consumption:
$100 + (0.70 \times 70) + (0.49 \times 70) + (0.343 \times 70) + \cdots = 263.3.$

Abstention:
$80 + (0.70 \times 80) + (0.49 \times 80) + (0.343 \times 80) + \cdots = 266.7.$

70% discounter starting off as consumer:

Consumption:
$70 + (0.70 \times 70) + (0.49 \times 70) + (0.343 \times 70) + \cdots = 233.3.$

Abstention:
$40 + (0.70 \times 80) + (0.49 \times 80) + (0.343 \times 80) + \cdots = 226.7.$

Perfect indifference arises with a discount rate of slightly more than 70%.

We can see from this that there is a range of discount rates somewhere above 60% and somewhere below 90% in which the choice to use or not depends on past usage.

The simplified B-M model is intended to show that addictive behavior can be rationalized: continued addiction can be seen expected utility-maximizing behavior. The model may be complicated to yield other aspects of addiction. For instance, one mark of addiction is that it may seem irrational to quit *today*, instead of tomorrow or in some future time-period. We can incorporate this aspect of addiction into the model by increasing reinforcement: by assigning greater disutility associated with symptoms of withdrawal. In this way, keeping up with the joneses may outweigh abstention even for those with a very low discount rate, such as a 90% discounter.

A puzzle about addiction

An interesting puzzle emerges from the simplified B-M model: *forward-looking* rationality (according to the principle of expected utility maximization) comes apart from a *backward-looking* assessment of utility (or welfare). It is rational for a 60% discounter to consume in all future periods, regardless of whether he consumed or abstained in the past. But at any point in time, we can see that *he would have been better off* if he had chosen to abstain *n*-periods ago, and continued to abstain through to the present day. By "he would have been better off" I mean both that he would be better off today (80 instead of 70), and that the sum total of his utility over the past *n*-periods would have been greater. The sum total of utility from abstaining *n*-periods ago and continuing to abstain through the present day is given by:

$$\sum_{n=1}^{n}(80_n)$$

This is greater than the sum total of utility he would have if he had chosen to consume *n*-periods ago, and continued to consume through the present day. It is given by:

$$100 + \sum_{n=1}^{n}(70_n)$$

No discounting is warranted here because there is no reason to discount past utilities (it may even be true that one's past utilities are *necessary* in the actual world).

But, if we do not discount past utilities, then of what value is the sum total of past utility, apart from derivative psychological effects, e.g., good memories? I think that past-utility

has an important role in our practical deliberations in that it allows us *to learn* from past mistakes. And this seems an important component in ones' rationally practical deliberations. It is an answer to the question "how *much better* things could have been?" And this seems a very important consideration in an agent's future practical deliberations. It is the stuff that *plans* are made in light of.

The upshot is that because it is rational for a 60% discounter to consume in any time-period, regardless of past consumption, it follows that it is *ir*rational for him to abstain in any time-period. He is, therefore, committed to the following claim: *I would have been better off if I had abstained n-periods ago, but it would have been irrational for me to do so*! This is a very peculiar result that, to be clear, does not arise because one's expected utility is lower than one's backward-looking assessment. This will always be the case with discounting. And the peculiarity does not arise because a 60% discounter's expected utility is lower than a 90% discounter's expected utility. The utilities are ordinal, and interpersonal comparisons are strictly forbidden (even *meaningless*, some have mistakenly thought in the past).

For a 90% discounter, on the other hand, the action recommended by the principle of expected utility maximization is consistent with his backward-looking assessment. It is rational for him to abstain in all time-periods. After *n*-periods of abstention, he may look back and conclude that he is better off (in both senses) than he would be if he had chosen to consume and continued to consume to the present-day. The result is mixed for the 70% discounter. It is rational for him abstain only if he abstained in the previous period. This will be fortunate if he finds himself in this position. If ever he consumes, rationality condemns him to a sub-optimal future.

Discounting and risk

The claim that a consumer's discount rate is a psychological constant, an exogenous variable that lies outside of the rational choice theorist's domain, is an entrenched assumption held by most in the economic community. Economic rationality treats discount rates as empirical inputs, and the rational choice theorist need not provide a normative basis for these rates. The same is true about attitudes toward risk: they are an empirically determined input, and rational decisions are made only in light of such attitudes. Providing a normative basis for discount rates and attitudes toward risk are no part of the traditional conception of economic rationality.

But advances in experimental economics have shown people to display a rich structure in their (broadly construed) economic decision making that is sensitive to slight changes in the description of a decision-problem, to changes in the quantity of the variables (e.g., the amounts of money they stand to win or lose), and to changes in the impact their decisions are expected to have on others. Some theorists conclude that this shows that people often act irrationally. No doubt this is the right conclusion to draw for some of these cases. But other theorists have taken away the idea that the received view of economic rationality needs to be revisited and made to fit at least some of these relatively new findings in experimental economics.

I believe that part of an improved conception of economic rationality is one that provides a normative basis for how much we *ought to* discount future utility, as well as for the types of attitudes toward risk that we should have. That attitudes toward risk should be thought of normatively—that they require justification, or must fit within constraints—can be made plausible with a simple thought-experiment. As if by miracle,

suppose a person finds himself with the choice between a guaranteed $4,000,000 and a fair coin flip for $0 (heads) or $10,000,000 (tails). Most people, I venture to say (not the Sultan of Brunei), would hasten to accept the $4,000,000, even though the expected (monetary) value of the coin toss is $5,000,000. And most people would surely accept a guaranteed amount of much less than $4,000,000 because most people *are* risk-averse. But extremely *risk-loving* people are rationally required to choose the coin flip.

Although I will not argue for the claim here, my swelling intuition is that it is not just reckless to choose the coin flip, but that it is irrational. Instead of saying that extremely risk-loving people are rationally required to choose the coin flip, I find it more plausible to say that people are rationally required not to be extremely risk-loving *in this choice situation* (all other things being equal). It is irrational to put one's future welfare at such risk (all other things being equal).

It is an interesting question whether a conceptual bridge can be constructed between attitudes toward risk and the degree to which one discounts future utility. My intuition expressed above leads me to think that there is one because the tentative support I offered for the claim that it is irrational to choose the coin flip over the guaranteed money is that it is wrong to take such great risks with one's future welfare. This suggests that extremely risk-loving people heavily discount their future. They are willing to take enormous risks only because they happen to care little about their future selves. If so, then the degree to which one discounts future utility has normative implications for a person's appetite for risk. Again, these are just suggestions, not arguments. At the very least, we can say that there does appear to be something odd, if not inconsistent, about an

agent who values his future welfare very highly, but who is also extremely risk-loving. The two do not appear to go together very well.

A principle of intrapersonal-prudential constraint

How, then, are we to find a normative basis for the discounting of future welfare? I offer only a limited answer to this question by proposing the following principle:

> *Principle of intrapersonal-prudential constraint*: prudential rationality requires of agents to refrain from acting on the principle of expected utility-maximization when it is inconsistent with a *projected* backward-looking assessment of their welfare (all other things being equal).

By "a *projected* backward-looking assessment" of welfare I mean an answer to the question "*would a consumer have been better off by choosing the alternative path*?" when answered from the perspective of one's future selves. As before, an answer to this question can take two forms: 1) better off in a future time-period (higher utility numbers on that day), and 2) the sum total of utility between one's choice and a future period is greater than the alternative when evaluated from the perspective of that person in a future time-period. I intend the principle to rule out (as irrational) action that is based on expected utility-maximization that cannot answer **both** forms of the question affirmatively. I leave open what conclusions we should draw from a mixed answer. Additionally, the principle only applies to cases that display a utility-structure like that in the simplified B-M model, where the utility from consumption is not expected to magically increase in the future. For example, someone may determine that he will be better off six months from now, and will have been better off six months from now (sum total of utility when evaluated from that future self), if had chosen a career in beach

volleyball, instead of enrolling in medical school. Furthermore, how far out into the future we must go until a consumer's backward-looking assessment of his welfare is inconsistent with his present-day calculation of his expected utility-maximization is sure to affect whether the principle is applicable. As Keynes famously stated: "in the long, we are all dead."

The implications for the simplified B-M model is that consumption in any period is irrational, given the assumption that the periods represent days, weeks, or months, and not decades or longer. From this, I take away the lesson that a 60% discounter ought to value his future welfare more. I claim this to be a requirement of prudential rationality. Given the numbers used above, a rational consumer ought to have a discount rate of somewhere in the low seventies.

Intuitively, I find the principle to be true because I do not think that it can be deemed rational an agent who determines that he would have been better off if he had abstained some relatively short time in the past and continued to abstain to the present-day, but goes ahead anyway and *consumes on* into the future. There is something important that he has not learned. I think this something is a requirement of prudential rationality.

Philosophically, I think that the principle of intrapersonal-prudential constraint is justified because many of the plans that we undertake presuppose it. As I said above, the reason we set plans is that they structure our lives. This allows us to pursue longer-term goals whose accomplishment is expected to make us better off. This expresses a basic commitment to pursue only those plans that are expected to make us better off. I allow that moral commitments may override the undertaking of some plans. But in the absence

of moral constraints, prudential rationality requires that an agent's plans our consistent with his conception of his well-being.