Meta-Preferences and Multiple Selves

By Douglas Glen Whitman
Dept. of Economics
California State University, Northridge
18111 Nordhoff St.
Northridge, CA 91330
glen.whitman@csun.edu
818-677-4542

THIRD DRAFT, completed 6 April 2004
DO NOT CITE

Meta-Preferences and Multiple Selves

ABSTRACT: A growing literature in economics and philosophy seeks to explain human choice in terms of either multiple selves or multiple layers of preferences, with the intent of explaining anomalies such as self-constraining behavior and time inconsistency. This article challenges the value of such models. While they do not pose unavoidable logical contradictions, they do pose analytical difficulties that cast doubt on their plausibility. They add superfluous layers of modeling complexity. They also err by trying to identify the "self" and its preferences with specific, narrow aspects of a person's thought. A simpler and more fruitful approach associates selfhood and preferences with the process by which the many aspects of a person's thought interact. This process will necessarily involve the dynamic discovery and creation of preferences over time.

Introduction

A growing literature in both economics and philosophy seeks to explain human choice in terms of either multiple selves or multiple layers of preferences.[1] The resulting models have been referred to as "egonomics,"[2] the economics of self-control,[3] the economics of self-government,[4] and so on.

Models of multiple selves seek to understand individual choice by treating it as analogous to social choice. The individual is actually not a unified self, but encompasses more than one self battling – or perhaps negotiating – for control of the individual's body and behavior. Among the possibly unsettling implications of this approach is that individual choice may be subject to the same problems and paradoxes that afflict social

[1] Citations of the relevant papers can be found in the appropriate sections below.
[2] Schelling (1978), Elster (1985), Klein (1992), et al.
[3] Thaler and Shefrin (1981), Elster (1985).
[4] Rubenfeld (2001).

choice, including the invalidity of interpersonal utility comparisons[5] and the possibility of intransitive choice orderings (i.e., Condorcet's paradox).

Models of multiple layers of preferences, also known as meta-preferences, understand human choice as resulting not just from preferences, but preferences about preferences, and perhaps even preferences about preferences about preferences… Meta-preference models take seriously the intuitively obvious fact that human beings pass judgment on their own desires and sometimes take action to change them. A troubling implication of meta-preference models is that individuals may have preferences they do not wish to have, and therefore may act in ways they do not wish to act.[6]

Though their approaches differ, meta-preferences and multiple selves models share common motivations and reach similar (though not identical) conclusions. In terms of motivation, both spring from the desire to get inside the black box notion of preferences employed in traditional economics. By so doing, they aspire to explain some seeming anomalies of choice, including: the tendency of humans to make statements that seem more consistent with multiple preference sets than with unified preferences; the prevalence of self-constraining behaviors, such as restricting one's access to desired substances (cigarettes, fattening food) or entering a rehabilitation clinic that restricts one's freedom of movement; and time-inconsistency, the seeming tendency of humans to privilege the present over the future simply because it is the present.

In terms of conclusions, both meta-preferences and multiple selves have the potential to drive a wedge between one's self and one's preferences, where preferences are understood as the desires that govern one's actual behavior. To put it another way,

[5] See Strotz (1955/6, 179): "The individual over time is an infinity of individuals, and the familiar problems of interpersonal utility comparisons are there to plague us."
[6] See, generally, George (1993, 1998, 2001a, 2001b).

both approaches invite us to distinguish between an individual's "welfare utility function" and his "behavioral utility function." While the latter is what determines the individual's behavior, the former determines his personal well-being. If these two utility functions are not the same, then individuals may systematically fail to act in their own best interests, even as they perceive those interests themselves.

In this article, I will challenge the value of meta-preferences and multiple selves models. Their motivating questions do deserve answers, and some of their conclusions have (limited) validity. But the models themselves add a layer of complexity to our understanding of human choice that is superfluous at best and misleading at worst. Both meta-preference and multiple-selves approaches err by associating the preference orderings of traditional economic theory with specific, narrow aspects of a person's thought. It is both simpler and more fruitful to associate selfhood and preferences with the process by which the many aspects of one's thought interact.

Sections I and II address meta-preferences and multiple selves respectively, defining them more precisely and exploring the difficulties associated with each. Section III presents an alternative analytical framework that overcomes the problems of both via a simpler and more plausible understanding of human choice. Finally, section IV examines the broader implications of the new analytical framework, which identifies both self and preferences with a dynamic process of discovery and creation, rather than a static ranking of situations.

I. Meta-Preferences

A meta-preference is a preference that one has about one's own preferences. In its most basic form, a meta-preference is simply a preference to have, or not have, another preference. Probably the most popular illustration of a meta-preference is a statement like the following:

(1) I want to smoke, but I wish I didn't want to smoke.

As we will see, even this simplest case is not clearly an example of a true meta-preference. Nonetheless, statements like (1) are so commonplace that they provide a useful touchstone in discussing what it means to have meta-preferences. (The distinction between desires and preferences will be addressed later.)

Meta-preferences are sometimes referred to as "second-order" preferences or, if we allow more than two layers, "higher-order" preferences. Such preferences are contrasted with the those preferences that do not make reference to any other preference, which are called "first-order" or "lower-order" preferences. I will use the terms meta-preference, second-order preference, and higher-order preference interchangeably, except when a distinction among more than two levels is required. I will also use first-order preference and lower-order preference interchangeably, with the same caveat.

A. Formal Soundness of Meta-preferences

The notion of a meta-preference is *formally* sound. It is simply a matter of defining wants or desires in such a way that a desire can itself be an object of desire. Frankfurt (1971) formalizes meta-preferences like so. The general form of any expression of desire is:

(2) A wants to X.

X is defined as an act or state of being; e.g., to consume a good or to engage in activity. If X is itself the act of wanting something, as in the following:

(3) X = want to P

then the desire expressed in (1) takes the form of a second-order preference:

(4) A wants to want to P.

Note that expression (4) does not imply A wants to P, as there is no logical inconsistency in saying both "A does not want to P" and "A wants to want to P."

George (1998, 2001a, 2001b) and Jeffrey (1974) offer formalizations in terms more familiar to an economist. Any preference takes the form

(5) X pref Y

or "X is preferred to Y." The objects being evaluated need not be simple things; they may themselves be statements of preference. Thus, if we have

(6) X = {P pref Q} and Y = {Q pref P}

then the idea expressed in (5) takes the form of a meta-preference:

(7) {P pref Q} pref {Q pref P}

The Frankfurt and George/Jeffrey approaches are similar, in that both involve the recursive application of some operator (desire or preference). The principal difference between the two approaches is that Frankfurt's approach does not imply that a stated desire is sufficient to trump other desires. An expression like (1) "A wants X" does not necessarily mean that A will do X, as conflicting desires (such as Y) may trump it. George and Jeffrey, on the other hand, use the notion of preference and the corresponding notation in the usual manner of economic theory. An expression like (4) "X pref Y" is taken to mean that if X and Y are both in the choice set, the agent will in fact choose X.

The George/Jeffrey approach may be extended to account for preferences over many different states of the world instead of just two at a time. A utility function (assuming the conditions necessary for one to exist) expresses a set of preferences over which an agent might conceivably have meta-preferences. Thus, if $u^1(x,y)$ and $u^2(x,y)$ are two utility functions over the goods x and y, then one might have the meta-preference

(8) $u^1(x,y)$ pref $u^2(x,y)$

which indicates that the first utility function is preferred to the second.

B. Comparability of Preferences and Utility

However, once preference is connected to utility, the concept of a meta-preference gets fuzzier. For if "A is preferred to B" is understood to mean "some agent gets greater satisfaction [in a sufficiently broad sense of that word] from A than from B," and if "utility" is understood as a metric for satisfaction, then it follows that some utility must come from the satisfaction of a meta-preference. A person who makes the statement in (1) must derive some satisfaction from not having a preference to smoke.

If utility can be generated through the fulfillment of both preference types, meta- or not, the natural question is whether the resulting utilities are different in some fundamental way. Utility is, of course, an imaginary construct, a metaphor for the way in which people compare and trade off different sorts of satisfaction, so we could write off this problem as a by-product of reifying utility. But if we set aside the notion of utility, the question re-emerges as the simple matter of whether a person can compare the satisfaction from the fulfillment of preferences at different levels. Can the happiness

generated by eating a candy bar be compared to the happiness generated by wanting to want a candy bar (or wanting *not* to want it)?

If the answer is yes, then a meta-preference can be effective in actual decisions. This point has not always been clear in the meta-preference literature. Sometimes the apparent position is that meta-preferences are never operative; in other cases, it seems that meta-preferences are operative only at special "moments" or under special circumstances. For example, George (1998) indicates that first-order preferences are always felt in action. If an agent has a first-order preference for smoking over non-smoking (S pref N), then he will necessarily choose to smoke, regardless of his meta-preferences: "Though preferring to have a preference to abstain from smoking, his first-order preference is to smoke and consequently he does smoke [if smoking is in the choice set]" (George 1998, 189). Similarly, Buchanan (1979b, 104) says, "the artifactual person he [an acting man] becomes does, at any moment, maximize utility subject to constraints." Nonetheless, meta-preferences can operate under specific circumstances, such as when an individual binds himself in advance: "[We can] allow for the possibility that one who is restricting one's future freedom is motivated by a second-order preference" (George 1998, 183); "But constraints on his own behavior imposed in past periods have shaped the form of his utility function, perhaps out of all recognition in some primitivistic sense" (Buchanan 1979b, 104).[7]

George's example of different levels of preference being operative at different times is instructive. On Wednesday, I predict that on Saturday – when I'm at a bar – I will wish to smoke. But my meta-preference for not wanting to smoke might cause me to

[7] Buchanan does not explicitly adopt a meta-preferences theory. Some aspects of this essay (1979b) are more consistent with the process approach I adopt in section III.

flush my cigarettes down the toilet now, in order to bind myself not to smoke on Saturday. Thus, the meta-preference is operative at a time and place (Wednesday at home) when the first-order preference is not.[8]

The problem with this reasoning is that the very same act might invoke both kinds of preference. As in the case of all addictive products, the present choice to smoke cigarettes affects both my present satisfaction and whether I will wish to smoke cigarettes in the future. So, suppose there are two time periods, 1 and 2, I am currently in period 1, and preferences are like so:

(9) S_1 pref N_1

(10) $\{N_2 \text{ pref } S_2\}$ pref $\{S_2 \text{ pref } N_2\}$

Thus, I wish to smoke now (9), and I also have a meta-preference not to wish to smoke in the future (10). The fulfillment or non-fulfillment of both preferences depends on my current (period 1) choice, because refraining from smoking in period 1 will help break my addiction and thus improve my ability to resist in period 2. Is there some reason to believe that one type of preference must always prevail over the other – i.e., that the smoker will always follow his present first-order preference rather than his present meta-preference about the future? If not – and the existence of many people who have kicked the habit supports this conclusion – then it seems inescapable that individuals compare these different sorts of satisfaction.[9] Preferences and meta-preferences cannot, therefore, be regarded as fundamentally incomparable.[10]

---

[8] Notably, George indicates that it would not be sensible to take this action unless it actually changed my future preference to non-smoking; otherwise, I would be denying my first-order preference to smoke while still failing to satisfy my meta-preference. "[W]ithin the analytical structure that I am proposing, the 'meddlesome self' increases his welfare only if the act of eliminating from the choice set that which he would prefer not to prefer succeeds in shifting his first-order preference" (George 1998, 189).

[9] In some places, George implies that it may be a present *first-order* preference about the future that motivates the self-binding action. See, e.g., Figure 4 in George 1998, 188. On Wednesday, the agent has

C. Truistic or Vacuous Meta-Preferences

Does the statement "I want to be satisfied," or "I want to be more satisfied," contain any idea that is not already contained in the very notion of "satisfaction"? I think not, unless satisfaction is described in a narrow fashion that excludes some types of desire, pleasure, or fulfillment. If satisfaction is understood in the inclusive sense usually employed by economists, statements like these are essentially vacuous, as are similar statements such as "I want to be happier" and "I want to have greater utility."

Many statements with the form of a meta-preference have this sort of vacuous content. They simply express a wish that one could be better off in some way, or they expression dissatisfaction with the state of the world and how it impinges on one's satisfaction.

To see this point more clearly, let us consider some simple cardinal utility functions:

(11) $u(x,y) = x^{1/4}y^{3/4}$

(12) $u(x,y) = 2x^{1/4}y^{3/4}$

(13) $u(x,y) = 2 + x^{1/4}y^{3/4}$

(14) $u(x,y) = x^{3/4}y^{1/4}$

---

both a second-order preference not to want to smoke on Saturday *and* a first-order preference not to smoke on Saturday. So perhaps it is only the latter that leads to the act of binding oneself. However, this approach doesn't escape the problem noted in the text. If the binding action is the act of abstaining from a cigarette now, then (in this version) I may simultaneously have a first-order preference to smoke (vis-à-vis my present satisfaction) and a first-order preference not to smoke (in order to affect my future preferences). And in George's approach, first-order preferences are "overall" preferences, so this is not just a mere conflict of desires – it is a logical contradiction.

[10] Jeffrey (1974) appears to see this point when he insists on the possibility of statements like the following: [{N pref S} pref {S pref N}] pref {S pref N}. In a statement like this, the agent shows a preference for the meta-preference over the preference. Jeffrey indicates that such statements can only make sense if the meaning of the "pref" operator is weakened: "[O]ne may have (and commonly, does have) various degrees of belief between 0 and 1 in sentences of the form 'A pref B' where 'pref' refers to one's own preferences" (Jeffrey 1974, 383). See discussion in section E, below.

It should be apparent that utility functions (12) and (13) are both superior to utility function (11), because they do not differ in any way from (11) except that they generate greater utility for any given level of consumption. It would be perfectly intelligible to say that (12) and (13) are preferred to (11), but it would also be vacuous: *of course* they are preferable, because more utility is by definition better than less. To say (12) and (12) are better than (11) is essentially to say, "I want to be more satisfied."

Now compare (12) and (13) to each other. Here, the choice is not as clear; (13) is superior to (12) if the individual expects to have relative low levels of x and y consumption, but (12) is superior to (13) if the individual expects to have relatively high levels of x and y consumption. In short, one's preference over (12) and (13) depends on a judgment about the state of the world. It would not be surprising for an individual who is poor (and expects to remain so) to say (13) is preferable to (12). It is like saying, "I expect to be poor, and I want to be more satisfied." But this is, like the preference for (12) over (11), a largely vacuous statement; the only additional information conveyed is a belief about the world, not the wants or desires of the individual.

For the utility functions considered so far, the consumption choices generated would be identical, as all three imply the same marginal rates of substitution (MRS) between goods. But now compare utility functions (11) and (14), which would generate different MRS's and thus, under most circumstances, different choices. If good x is expected to be relatively easier to obtain (e.g., have a lower price) than good y, (14) is the superior utility function; if good y is expected to be relatively easier to obtain, (11) is superior. If an individual indicated that (14) is preferable to (11), we could interpret him as saying, "I expect that good y will be relatively cheaper than x, and I want to be more

satisfied." Again, the statement is mostly vacuous; the only non-truistic information conveyed is a belief about the world – this time, about the relative prices of goods.

In comparing the first three utility functions, we can be relatively confident about their vacuity. The difference in total utility is the *only* difference between them. But the comparison of (11) and (14) is open to other interpretations. One possibility is that the individual thinks, for some reason, a desire for y is inherently better than a desire for x, regardless of prices. Perhaps we could interrogate our subject: Would you prefer (14) to (11) regardless of the relative prices of the goods?[11] If the answer were yes (and truthful!), then we might infer the existence of true meta-preferences, that is, preferences conveying more information about the wants of an individual than a mere desire for more utility. (Or perhaps not – we will consider other explanations later.)

The point here is not that *all* meta-preferences are vacuous, only that *many* such statements may indeed be vacuous, even if exceptions exist. Take, for instance, a statement that I have made on many occasions: "I wish I enjoyed exercising more." This could be taken as a preference over preferences: I want a utility function with a higher MRS of exercise for other activities. But the only reason I make this statement is that I happen to live in a world in which exercise is desirable for a variety of reasons such as health, attractiveness to other people, and so on. We could imagine a different world – one with tastier health food and more efficient muscle-building technologies – in which I would not express such a desire. Although we can imagine a person who believes that exercise is inherently superior to relaxation, I can assure the reader that I am not that person! My supposed meta-preference turns out to be a simple wish to have more utility.

[11] Instead of insisting on a preference for (13) *regardless* of prices, we could look for a preference for (13) that persists even in the face of a very high (but not infinitely high) relative price of good x.

Traditionally, microeconomic theory sharply distinguishes preferences from constraints. Consumer choice results from the interaction of preferences and constraints, which can be defined separately and without reference to each other. Consumer preferences can be stated without reference to prices and income, for instance. But some alleged meta-preferences apparently result from transgressing the preference-constraint dichotomy. My meta-preference for exercise over relaxation derives from my recognition of constraints imposed by the world I live in.

Crossing the preference-constraint boundary is not necessarily a bad thing; it happens sometimes even within traditional microeconomics. The most obvious example is the characteristics of a single good, which are regarded as given when preferences over different goods are defined. Take the individual who says he prefers oranges to apples:

(15) orange pref apple

This sounds like a pure statement of preferences, until we realize that his more complete preferences are like so:

(16) (apple without seeds) pref (orange) pref (apple with seeds)

This individual's preference for oranges over apples actually results from the incorporation of a constraint: that apples always have seeds, or that it is costly to seed one's apples. We would not say that the preference in (15) fulfills a meta-preference or that its negation denies one; instead, we would say that (15) summarizes the interaction of more basic preferences and constraints taken as given (for the time being). Similarly, an individual's stated meta-preference for non-smoking over smoking might represent his recognition of a constraint that eliminates healthful smoking from his choice set.

D. Meta-Preferences as Intellectualized Desires

The term preference, as used by economists, differs somewhat from its vernacular use. It is common to hear statements like

(17) I would prefer to sleep in, but I get out of bed and go to work instead.

And this would be a perfectly acceptable use of the word "prefer" to most people. But we would not generally interpret (17) to mean the agent deliberately acted counter to his own preferences in the broadest sense. Instead, the "preference" to sleep is understood as *one* desire the agent has, considered in isolation from other desires. In general, economists use "preference" in a more inclusive sense, so that a statement like (4) "X pref Y" means that X is preferred over Y *all things considered*.

Henceforth, I will use the term "desire" to indicate a source of satisfaction considered apart from others, and I will use "preference" in the usual economists' sense of an all-inclusive statement of the superiority of one option over another. (Baier [1977] and George [1998] use the terms "intrinsic preference" and "overall preference" to track the same distinction.)

In general, a preference statement indicates the relative levels of satisfaction or dissatisfaction attributable to the objects that it ranks. The statement X pref Y indicates that the having (or doing) of X generates a greater level of satisfaction for the agent than does Y. This is true even for meta-preferences. The statement in (7), for instance, indicates that the having of preference {P pref Q} generates a greater level of satisfaction for the agent than does the having of preference {Q pref P}.

For a meta-preference to be truly meta-, then, the objects of satisfaction it refers to cannot be the same as those of a first-order preference. A meta-preference is not simply a

"superior ranking" of the elements ranked by a first-order preference; if it were, it could be stated more simply as a first-order preference.[12] For the statement in (7) to contain any additional information about the agent's wants, it must *not* be equivalent to the statement {P pref Q}. Alternatively, if we choose to regard a meta-preference as another (possibly superior) ranking of the same elements as first-order preferences, then we have transformed the meta-preference approach into a multiple-selves approach. The merits and demerits of that approach will be considered separately in part 2.

It seems likely, though, that many apparent statements of meta-preference are not truly meta-preferential in the sense just discussed. They do not indicate satisfaction from the having of certain first-order preferences *per se*, but from the actions or states of being that are the objects of those first-order preferences. Take for instance, the claim in (1), "I want to smoke, but I wish I didn't want to smoke." It seems sensible to characterize this person's preference structure like so:

(18) S pref N

(19) {N pref S} pref {S pref N}

This a familiar situation, and it becomes even more familiar when extended to other goods and activities with pleasant short-term effects and unpleasant long-term effects, such as drug use and consumption of fatty foods. Most of us have experienced the kind of tension captured by (18) and (19) at one time or another. I suspect it is this familiarity that makes the notion of meta-preferences attractive to many thinkers.

But *why* might one have meta-preferences like those expressed in (19)? In the case of smoking, the answer seems obvious: smoking harms one's health, shortens one's

[12] See George's (1993, 329) discussion of Hollis (1983), who does conceive of meta-preferences as a superior ranking of the same elements.

life expectancy, poses a fire hazard, makes one's clothes smell bad, etc. But these are all desires about *outcomes*, fully expressible as simple first-order preferences. The fact that smoking dominates non-smoking does not mean that non-smoking is valueless. The alleged statement of meta-preference in (18) may thus be nothing more than the amplification of one desire in one's first-order preferences, or one argument in one's standard utility function.

Put another way, some alleged second-order preferences are just intellectualized expressions of first-order desires.[13] Humans, unlike other animals, have the capacity for self-reflection; they can not only have a desire, but recognize it and even pass judgment on it. It is this capacity that motivates Frankfurt (1971) to identify the possession of second-order desires[14] as the (or a) unique attribute of personhood. But the act of recognizing a desire does not necessarily make that desire more genuine than any other. This becomes apparent if we ask what motivated an act of self-reflection, and we find that it was a first-order desire – typically a frustrated one. The continuing presence of unsatisfied desires, which follows from living in a world of scarcity and other constraints, demands psychic attention: the agent examines his unsatisfied desires, asking whether they are really compensated by those desires that are satisfied.

That meta-preferences are often the manifestation of frustrated first-order desires is evidenced by the virtual absence of meta-preferential statements of satisfaction. We hear, "That apple pie really hit the spot" (an expression of first-order preference

[13] Previous authors, especially George (2001, 325), sharply distinguish dissatisfaction due to unmet meta-preferences from dissatisfaction due to unmet first-order desires. My point here is that, despite the formal distinction, the appeal of meta-preference models may be attributable in part to a subtle conflation of the two.

[14] Actually, second-order volitions, which are second-order desires that an agent would wish to make effective in action. Frankfurt (1971), 10.

satisfaction); we do *not* hear, "I sure am glad I like apple pie" (an expression of second-order preference satisfaction). Statements of alleged meta-preference seem to refer almost exclusively to sources of *dis*satisfaction.

Not all first-order desires are equally susceptible to intellectualization. The desires most often cloaked as meta-preferences are, not surprisingly, those with greater social approbation. In modern times, smoking is no longer a politically correct lifestyle choice. An even better example of the role of social approval in the formation of alleged meta-preferences is provided by Friedman (1986, 31), who describes a woman who has been taught to believe that "a woman's place is in the home." This attitude can be expressed as a meta-preference: she would prefer to prefer homemaking to careerism.

George (1998) observes that unfulfilled meta-preferences can be a source of dissatisfaction. The agent who prefers A to B, but who would prefer to prefer B to A, experiences a sort of disharmony. In Friedman's example, a woman who has chosen careerism over homemaking might experience feelings of ongoing regret. A meta-preference theorist might attribute her feelings of regret to the disharmony between her preference (she prefers careerism) and her meta-preference (she prefers to prefer homemaking). But what purpose is served by the additional structure? Her feelings of regret could just as easily be attributed to her unfulfilled desires for homemaking, which are outweighed by her desire for careerism but persist nonetheless. Left unfulfilled, these desires are intellectualized in the form of a so-called meta-preference.

To amplify the point, imagine an alternative scenario in which the woman chooses, on the basis of her meta-preference for homemaking, to create obligations that bind her to a life in the home. Perhaps her binding activities actually succeed in altering

her first-order preferences; e.g., having children creates in her a greater devotion to the value of motherhood, while her lack of education creates a fear of failure in the workplace. Yet she continues to experience "frustration, guilt, and depression" (31) because of her latent desire for a career outside the home. We do not require meta-preferences to explain these feelings. Indeed, meta-preferences *cannot* explain them, because her preferences and meta-preferences are consistent: she wants what she wants to want. So if some feelings of regret (from not pursuing a career) are attributed to unfulfilled desires already included in her first-order preferences, why attribute other feelings of regret (from not pursuing homemaking) to unfulfilled meta-preferences?

E. True Meta-Preferences?

The preceding discussion suggests that, at least much of the time, what seem to be meta-preferences are really just expressions of unfulfilled desires. Now we will attempt to find "true" meta-preferences. The smoking gun would be a meta-preference whose fulfillment by acquisition of the appropriate first-order preference would generate some satisfaction *even if the first-order preference were not fulfilled*.

Suppose, for instance, that someone with the preference structure of (18) and (19) – someone who prefers to smoke and prefers to prefer not to smoke – could, at zero cost, alter his first-order preference (18) so that he now prefers non-smoking. The catch is that he would be forced to smoke anyway, and thus his new first-order preference would be unfulfilled. Would this person nonetheless gain some utility from simply having the new preference?[15] If so, then we would have a genuine meta-preference. The problem, of

[15] The satisfaction gained need not be large enough to overcome the dissatisfaction from being forced to smoke.

course, is that this seems rather unlikely; typically the person who wants not to be smoker wants it because of the health benefits from not smoking. If he were forced to smoke anyway, he would not get those benefits.

Frankfurt provides another possible example of a meta-preference whose fulfillment generates satisfaction even without the fulfillment of a first-order preference:

> Suppose that a physician engaged in psychotherapy with narcotics addicts believes that his ability to help his patients would be enhanced if he understood better what it is like for them to desire the drug to which they are addicted. Suppose that he is led in this way to want to have a desire for the drug. … It is entirely possible, however, that, although he wants to be moved by a desire to take the drug, he does not want this desire to be effective. He may not want it to move him all the way to action. (Frankfurt 1971, 9)

This example seems to meet the criterion set out above. The physician would get some satisfaction from having different first-order preferences even if those first-order preferences were not fulfilled. The possession of an unmet craving for drugs would create some amount of dissatisfaction, but the physician would find it to be outweighed by the satisfaction of having the new preference.

Even here, however, difficulties arise. First, why does the physician have this meta-preference? He wishes to serve his patients better, and that is a desire already included within his first-order preferences. The physician's dilemma could be cast as the tension between his first-order desire to help his patients and his first-order desire *not* to take narcotics, combined with a constraint that requires a trade-off between the two (specifically, that empathy requires common experience). Even if we recognize the physician's desire to desire narcotics as a true meta-preference, we still find a relationship between the meta-preference and some desire encompassed by his first-order preferences.

Consider another example. A devout Christian expresses a desire not to be greedy. Now, greed is itself a kind of preference (to accumulate wealth), and thus the Christian's desire not to be greedy has the form of a meta-preference. He prefers wealth to poverty, but he would prefer to prefer poverty to wealth.[16] Note that he could obtain satisfaction from the fulfillment of his meta-preference without the resulting first-order preference being fulfilled; if he happened to obtain wealth by accident, that would not mean he was greedy. And this case seems stronger than that of the physician, because the physician's meta-preference was rooted in a first-order preference to help patients, not in any belief that addiction is good as such. The Christian, on the other hand, obtains satisfaction strictly from the absence of greed; it is not the wealth itself that God condemns, but the *desire* for wealth.

Still, the Christian example might be susceptible to the same objection as the physician example. Whence comes this meta-preference for preferring poverty to wealth? The source is presumably a first-order desire to please God. We could cast the Christian's conflict as one between his first-order desire for wealth and his first-order desire to please God, combined with an assessment of what God wants. But in this case, what God wants from the agent is not an action vis-à-vis the outside world, but a mental state in the agent's head.

The physician and Christian examples provide (to this author's mind, at least) the most persuasive case for the existence of true meta-preferences. Still, the meta-preferences in both cases appear to be dependent on some first-order desire. They appear

[16] Alternatively, if the Christian opposition to greed does not imply support for poverty, he might prefer to be indifferent between wealth and poverty.

as *instruments* for the better achievement of first-order preferences.[17] Their importance is contingent on the importance of the lower-order desires they serve. A person might be willing to trade off some degree of fulfillment of one first-order desire in order to acquire a different preference ordering motivated by another first-order desire. For example, the physician might willingly sacrifice some amount of physical comfort – that which derives from not having a nagging craving for heroin – for the greater satisfaction he obtains from serving his patients better. An especially dedicated physician would presumably be willing to sacrifice more physical comfort than would a less dedicated physician.

The broader point is that, even if we accept meta-preferences whose origin is in first-order preferences as genuine, there is still a potential trade-off between different levels of preference. That means they are not fundamentally incomparable.[18] Moreover, since the agent could make the trade-off in either direction, neither level of preferences can be taken as the true source of an agent's actions. A full explanation of agent actions requires reference not just to first-order preferences, or even to first-order in some circumstances and meta-preferences in others, but to the interaction between the two.

II. Multiple Selves

The multiple selves approach to modeling internal conflict treats an individual as having more than one self, or more than one personality, whose interests may conflict. In other words, the different selves have different preferences. As multiple selves models

[17] See Weisbrod (1977), which says one utility function is preferable to another if the choices it generates do a better job of satisfying the other utility function.
[18] Jeffrey (1974) appears to recognize this point. See note 10, above.

are often deployed for the same purpose as meta-preference models, the paradigmatic example will look familiar:

(20) The fun-loving me wants to smoke, but the health-conscious me wants to stop.

Multiple selves appear in the vernacular, not just the literature of economics and philosophy. It is not uncommon to hear statements like these:

(21) One part of me wants to stay up late with my friends, but another part of me wants to go to bed so I can work in the morning.

(22) My rational mind tells me this relationship won't work, but my romantic mind tells me to pursue it anyway.

The multiple selves literature is inconsistent on the question of whether one self is somehow superior to the other. Sometimes there is no implication of superiority, only conflict; Cowen (1991) is an excellent example. In other cases, the language itself implies a superiority-inferiority relationship, as in (22), where "rational" is (at least to some) a judgment of correctness. In yet other cases, the analysis explicitly favors one self over other, as in Thaler and Shefrin's (1981) model, which pits a far-sighted "planning" self against many myopic "doer" selves; the former self is self-evidently superior. All will be considered here, although we will see that some approaches blur the line between multiple-selves and meta-preference models.

A. Formal Soundness of Multiple Selves

Multiple selves, like meta-preferences, is formally sound as a concept. Just as we can observe different individuals whose interests conflict with respect to the use of

external resources such as land, we can imagine different individuals whose interests conflict with respect to the use of a human body. And just as the former may fight for control over external resources, the latter may fight for control of the body. Formally, all we need do is posit two different preference sets or two different utility functions. In terms of desires, we may say:

(23) A_1 wants X.

(24) A_2 wants Y.

In terms of preference relations, we may say:

(25) X $pref_1$ Y

(26) Y $pref_2$ X

where the subscripts indicate different selves. And when we wish to cover more than binary choices, we can use utility functions with similar superscripts, such as $u^1(x,y)$ and $u^2(x,y)$. (We do not, however, indicate any preference between them as in (8).)

With multiple selves, as with meta-preferences, the devil is not primarily in the formal structure, but in the interpretation. The difficulties afflicting the former turn out to be similar, though not identical, to those afflicting the latter.

B. Comparability of Preferences and Utility

The first question is when, if ever, any specific self is in control, and thus when, if ever, its corresponding preferences are operative. To have explanatory power, the multiple selves model must define some kind of division of power or allocation of personal resources. A typical approach would say that one self takes command during certain times and in certain circumstances, while a different self commands at other times

and in other circumstances. Take the case of the smoker, with preferences like those in (20). We could suppose his short-term self will wish to smoke on Saturday, but his long-term self (looking forward from Wednesday) does not wish him to smoke on Saturday. The long-term self might therefore launch a preemptive strike by constraining the short-term self's ability to obtain cigarettes on Saturday. This is the approach adopted by Schelling (1996) and criticized by George (1998).

The difficulty here, analogous to the earlier criticism of meta-preferences in the same situation, arises when the very same action invokes both selves and thus both sets of preferences. In the example as presented, the long-term self is able to act as it pleases on Wednesday, because the impulse to smoke will not arise until Saturday. But one thing that could affect the likely action of the agent on Saturday is the choice of whether to smoke a cigarette on Wednesday; resisting the urge to smoke now can help break the habit and reduce the impulse later. If we regard the long-term self as having control over long-term decisions and the short-term self over short-term decisions, then both must have some control over a decision with both short- and long-term consequences. Is there some reason to believe that one self will always prevail? At the very least, a more precise definition of the balance of power or resource allocation is required.

A natural solution would be to compare the utilities of the two selves and ask which is greater in any given situation. That, however, is what the multiple-selves approach invites us *not* to do. The conventional wisdom in microeconomic theory is that intrapersonal utility comparisons are invalid (or meaningless), and the same conclusion arguably applies to the utility of different selves in the same person. If it does not, meaning the utilities of the selves can be compared and traded off against each other, then

the usefulness of the multiple-selves construct falls into doubt. The *true* utility function of the person with multiple selves, with utility functions $u^1(x,y)$ and $u^2(x,y)$, would presumably be something like $w(x,y) = \alpha u^1(x,y) + u^2(x,y)$, where α indicates the weight of the first self's interests relative to the second's. (A more complicated function would apply if there is not a linear trade-off between the two.)

C. Multiple Selves as Alienated Desires

Multiple selves might be nothing more than anthropomorphized desires, where desires are understood as the different sources of satisfaction within one's preferences. They correspond to arguments within one's utility function, not to an entire utility function. As evidence for this proposition, notice how the identified selves are typically one-faceted creatures who rarely, if ever, exhibit any kind of balance. The speaker of (20) identifies two selves, the health-conscious self and the fun-loving self. The speaker of (21) identifies two selves, the conscientious worker and the late-night partier. The speaker of (22) identifies two selves, the rational self and the romantic self.

Why not the somewhat-more-health-conscious-self and the somewhat-more-fun-loving self? The self who likes to party a lot and the self who likes to party a little? The mostly romantic self and mostly rational self? The selves-within-ourselves resemble archetypes more than real people. Their utility functions seem not to be $u^1(x,y)$ and $u^2(x,y)$, but $u^1(x)$ and $u^2(y)$.

The most plausible explanation for the strange one-sidedness of most identified multiple selves is that they are not, in fact, different selves with different preference sets, but aspects of a single self. In this sense, they are analogous to those meta-preferences

that grew from frustrated first-order desires; but instead of (necessarily) treating them as having a supervening relationship to the rest of our preferences, the person alienates them to a greater or lesser degree.

Alienation sometimes serves merely to categorize the conflicting sources of one's impulses. In other cases, the agent may alienate some desires more and identify with those that remain. To the extent that the agent identifies with some desires more than others, some selves may appear superior to others – hence the tendency of some multiple selves accounts to privilege some selves. It is in these models that the difference between meta-preferences and multiple selves is fuzziest. This is especially so in meta-preference accounts that treat meta-preferences as superior rankings of the same elements as first-order preferences.

The perception that some selves are necessarily superior to others should be guarded against. The desires most likely to be reified as superior selves, like the first-order desires most likely to be elevated to the status of meta-preferences, tend to be those with the greatest social approval. Friedman's case of a woman torn between her desire for a career and her belief that a woman's place is in the home could be cast as a conflict between multiple selves – homemaker self versus careerist self – as easily as a conflict between different levels of preference. She might alienate one of these selves more than the other, perhaps viewing it as more "external" to herself, though it is not a given which self would have that status.

Cowen (1991) makes a related point when he observes that the rule-oriented self is not necessarily any more "rational" than the impulsive self. (Cowen adopts this

terminology to distinguish his approach from the more typical long-run self versus short-run self.)

> There is no prima facie reason for believing that victory for the rule-oriented self is desirable. Many cases, such as the abuse of alcohol or hard drugs, may require the victory of the rule-oriented self for personal welfare, but too many victories for the rule-oriented self can be injurious to mental health. A person who continually thwarts the desires of his impulsive self may become frustrated and overly rigid and lose his capacity for spontaneity. (Cowen 1991, 365)

As possible instances of individuals who err on the side of too little spontaneity, Cowen offers tightwads, workaholics, and compulsive exercisers, among others (367).[19]

It is a truism that individuals face multiple and often conflicting desires. The traditional approach, with its unitary preference sets and utility functions, already captures that basic truth. To the extent that the multiple selves approach merely reifies impulses within a person as separate persons, it adds nothing useful to our perspective.

III. Self as Process: An Alternative to Meta-Preferences and Multiple Selves

The arguments in the previous two sections do not definitively dispose of meta-preference and multiple selves. Both approaches are formally sound, inasmuch as they do not pose any unavoidable logical contradictions. My intent was to cast doubt on their reality and usefulness. In this section, I take a more constructive tack by describing a distinct notion of selfhood, and corresponding notion of preferences, that avoids some of the difficulties afflicting the meta-preference and multiple-selves approaches.

[19] Although Cowen's model avoids the bias of those models that privilege long-run thinking and self-control, it still qualifies as a multiple-selves model and is subject to some of the same criticisms. Specifically, Cowen's rule-oriented and impulsive selves are susceptible to the criticism of being single-faceted: "Personality features such as self-control or spontaneity are linked to particular internal values. By exercising self-control, for instance, a person may better achieve the internal values of prudence and moderation. Or spontaneity may favor the value of sexuality" (Cowen 1991, 362).

A. Self as Process

Meta-preferences and multiple selves are attempts to model a big truth we all know intuitively: that human beings have various impulses that compete against each other. As I hope the earlier sections have established, perceived meta-preferences and multiple selves often arise from regular (first-order) desires of the individual. Models that employ meta-preferences and multiple selves are motivated, in large part, by the recognition of our competing desires. But that motivation is *already* built into the notion of a utility function or preference ordering. The purpose of a utility function or preference ordering is to represent a process by which competing desires and impulses are more or less rationalized. Adding more structure, in the form of competing utility functions or rankings of preference orderings, needlessly complicates the analysis.

Both approaches commit the error of thinking that utility functions and preference orderings, as traditionally used in economics, must be associated with any specific aspect of a person's thought. The viewpoint I adopt here is that the self, and the preferences corresponding to it, emerges through an internal *process of interaction* among the various aspects of one's thought, including desires that often conflict. By "self," I mean the unique personal qualities that define us as distinct human beings. The claim here is that what we identify as our unique personal qualities has much to do with how we resolve our internal conflicts.

Personification occasionally plays a role in the process, leading to an internal perception of what seem to be separate persons or selves. Sometimes we feel as though they are equal selves; other times we feel like one self is somehow superior to the others,

because it is more foresighted or more capable of adopting a seemingly disinterested perspective. But to characterize any one of these selves as the "true" self and thus the "true" source of preferences is an error. The self is not any one of these personified desires, but the interaction among them.

The viewpoint just stated is summarized in Figures 1 and 2 (see page 40). In both figures, *u(x)* is used as shorthand for preferences or a preference ordering, but should not be taken to imply all the necessary conditions for the existence of a cardinal utility representation of preferences. Figure 1 contrasts the present view of meta-preferences with the typical[20] view in the meta-preferences literature. The received view drives a wedge between the self and *u(x)* by identifying the former with one's highest order desires, the latter with one's lowest order desires. The present view identifies both self and *u(x)* with the process of interaction among the desires at different levels.

Figure 2 contrasts the present view of multiple selves with the typical view in the multiple-selves literature. The received view drives a wedge between self and *u(x)* by identifying a distinct self (with corresponding utility function) with each distinct set of desires. The presence of more than one set of desires implies a conflict between each self and the preferences associated with the other self. The present view, by contrast, identifies both self and *u(x)* with the process of interaction among different desire sets.

B. What Is to Be Explained?

[20] The characterization of meta-preferences here does not apply to all authors. For instance, George (1998) eschews the identification of the individual with his highest order preferences.

Those who employ meta-preferences and multiple-selves approaches assuredly wish to explain or accomplish something, so the natural question is whether the framework outlined here does the job.

Conflicting desires and unwanted constraints. My primary argument has been that many explananda of meta-preferences and multiple selves do not, in fact, require any further explanation. People have conflicting desires. Sometimes they are hard to rationalize them. Sometimes the fulfillment of desires is frustrated by constraints. Some desires are more easily fulfilled than others, and which desires are most easily fulfilled depends in part on the constraints faced. All of these observations are true, but not in need of further explanation.

Self-constraining behavior. Other items to be explained by meta-preferences and multiple selves are more difficult to rationalize in terms of mainstream theory. Specifically, it is difficult to explain the willingness of people to place constraints on themselves, at least when such constraints are not necessary to induce cooperation from others in a social context (as when one signs a contract). Why would a person deliberately reduce the size of his choice set and even pay to do so? Multiple selves and meta-preferences are intuitively appealing explanatory tools in these cases, because we all understand why people would want to place constraints on others. We "explain" the behavior by saying, in essence, that the individual is not really constraining himself, but rather constraining another person who will inhabit his body.

Self-constraining behavior does pose a problem for the standard theory, but invoking the structure of meta-preferences or multiple selves is theoretical overkill. A more parsimonious explanation is simply that *people sometimes act irrationally*. That is,

under some circumstances people act in a manner that is inconsistent with their considered preferences. In terms of the framework presented above (and represented in the right-hand side of Figures 1 and 2), the utility function that results from the interaction of one's desires does not always issue in action. Instead, a single desire leads directly to action instead of going through the indirect process of interaction with other desires. Perhaps this explanation seems too simplistic. But what is added to our understanding by saying, instead, that a separate self "takes over" during certain circumstances? The "self" in question typically turns out to be strangely one-dimensional, concerned only about the gratification of one desire – in which case the one-desire explanation is both more plausible and more direct.

Although this position constitutes a departure from the "rationality-always" assumption, it need not imply a blanket rejection of rational choice modeling. It simply means that there are limited circumstances under which the rationality assumption does not apply. Moreover, rationality can still play a role in explaining how people deal with the fact that under some circumstances they will not be rational. Knowing that one might smoke a cigarette if it is available, the smoker trying to quit flushes the pack. Doing so is a rational method of coping with a predictable form of irrationality. This approach is equivalent to Schelling's "anticipated irrationality" (Schelling 1996) and Elster's notion of "second-best rationality" (Elster 1979)[21]. Notably, both Schelling and Elster have invoked multiple selves in their work, but multiple selves are unnecessary for the concepts just mentioned.

[21] "Man often is not rational, and rather exhibits *weakness of will*. Even when not rational, man knows that his is irrational and can *bind himself* to protect himself against the irrationality. This second-best or imperfect rationality takes care of both reason and passion" (Elster 1979, 111).

Time inconsistency. Many of the phenomena that meta-preferences and multiple selves purport to explain relate to delayed gratification, or the experience of costs and benefits at different points in time. Such situations have been dealt with extensively in the literature on time inconsistency.[22] The classic example of time inconsistency is an individual who, in an experimental setting, chooses a larger amount of money to be received in 101 days over a smaller amount of money to be received in 100 days – but who chooses the smaller amount of money when it will be received immediately instead of tomorrow. The existence of preference reversals like this is sometimes taken to show the existence of hyperbolic (or quasi-hyperbolic) preferences[23].

Do meta-preferences or multiple selves offer a better way of conceptualizing the problem of time inconsistency than do hyperbolic preferences? Or, alternatively, do hyperbolic preferences constitute a special case of meta-preferences or multiple selves? In terms of the framework presented here, the issue is again a matter of conflicting desires. I want to consume more goods *now*; I also want to consume more (perhaps different) goods *later*. I want to enjoy a cigarette *now*; I also want to feel healthy *later*. The mere fact that desires conflict does not justify adopting the more complex structure of meta-preferences or multiple selves. If some costs and benefits experienced in the present assert themselves more prominently than would be consistent with standard (exponential) discounting of the future[24], then we may have another case of *desires* rather than preferences issuing in action. The kinship between hyperbolic preferences and

[22] See, among others, Strotz (1955/6), Laibson (1997), Mulligan (1996).
[23] Hyperbolic preferences place greater weight on utility received in a given time period, relative to more distant time periods, the closer the period in question is to the present. They are contrasted with exponential preferences, which discount more distant periods relative to closer periods at a constant rate. Quasi-hyperbolic preferences are similar to hyperbolic preferences, except that the additional weight is attached only to the present period.
[24] I do not commit myself on the question of whether the evidence for hyperbolic preferences is strong enough to affirm their existence.

anticipated irrationality is illustrated by the fact that an individual with hyperbolic preferences can improve his expected utility by *binding himself* in advance to (for example) contribute more money to a savings account. In principle, this is no different from the forms of self-constraining behavior discussed above.

The position adopted here is consistent with Gifford's (2002) biology-based model of choice. Gifford observes that, for reasons relating to our evolutionary history, humans have a systematic tendency to select less abstract options over more abstract options. The future is necessarily more abstract than the present, which means humans often exhibit a tendency to weight present costs and benefits heavily relative to the future. But, Gifford notes, favoring the present over the future is only one instance of a more general phenomenon:

> Similar problems, however, can arise when making choices between two goods when both are available to the agent with a predetermined identical short delay. If one of the two goods is represented only by a printed word, for example, and the other good is visible to the agent when making the choice, reversing the level of abstraction of the two goods can result in a reversal of the agent's choice. (Gifford 2002, 114).

Gifford's approach offers a means of classifying the circumstances under which desires, rather than preferences, have a more pronounced tendency to result in action. He emphasizes the role of inhibition in creating a buffer zone in which *deliberation* can take place (Gifford 2002, 118-119). Deliberation is analogous to the process of negotiation depicted in Figures 1 and 2.

Desire-affecting choices. One last phenomenon that meta-preference and multiple-selves theorists might wish to explain is that people often take actions they know will change their desires in some way. They consume substances that could result in addictions, and then (sometimes) they try to break their addictions. They listen to a

form of music, knowing that doing so may affect their future ability to appreciate that form of music. With choices among goods, preferences over goods are sufficient; but with choices among desires, it might seem that preferences over desires are required.

Becker and Murphy (1988) have provided a consistent account of desire-affecting choices with their model of rational addiction. Their model does not require multiple selves or different layers of utility functions, but instead relies on a single utility function whose contents are affected systematically by prior consumption choices. Now, this approach arguably *does* incorporate multiple possible utility functions, inasmuch as the "single" utility function can take many forms depending on the agent's choices. The agent is choosing what instantaneous utility function to have in future periods. But on what basis is the agent making the choice? No meta-preferences need be invoked, except in the truistic sense discussed in section 1: the agent wants to be happier. Given his constraints, including the technical relationship between present consumption and the future form of his instantaneous utility function, the agent simply chooses that utility function (or stream of future utility functions) which creates the greatest utility. This is conceptually no different from choosing utility function (14) over utility function (11) when good x is expected to be relatively cheap.

There are, of course, objections to the rational addiction model, but the most salient criticism is simply that agents are less foresighted than the model implies. If true, this criticism could be incorporated via the use of a hyperbolic intertemporal utility function in the Becker-Murphy model; this is the strategy followed by Gruber and Koszegi (2000). As discussed earlier, hyperbolic preferences need not indicate the presence of multiple selves or meta-preferences.

IV. Creation and Discovery of Preferences

Although I have defended the traditional understanding of preferences against the challenge of meta-preferences and multiple selves, the traditional approach is not immune from criticism. Indeed, the framework sketched here shows how the traditional approach could be modified to better explain human behavior. The first modification, cited earlier, is the weakening of the "rationality always" assumption, and hence the notion of revealed preference. For most neoclassical economists, this would constitute a major departure.

The second modification is perhaps even more radical. If both self and preferences are to be identified with a *process*, as I have argued in this paper, the natural question is whether that process is ever complete. The answer, I posit, must be no. The process of negotiating among one's desires in order to find the optimal balance involves learning about both the world and oneself, and learning is always incomplete. New discoveries about the capacity of goods and actions to satisfy one's desires lead to new behaviors, and new behaviors lead to yet more discoveries. There is no reason to assume that the learning process has an endpoint, especially in a world wherein the characteristics of goods and the opportunities available also change continuously.

The notion that preferences reflect an ongoing discovery process has been hinted at by various authors working in the Austrian economic tradition. Although Israel Kirzner's theory of entrepreneurship is most commonly applied to the efforts of businesspeople to discover better ways of satisfying the existing wants of consumers, Kirzner emphasizes that there is an entrepreneurial aspect in all decision-making:

> Now I choose … to label that element of alertness to possibly newly worthwhile goals and to possibly newly available resources – which we have seen is absent from the notion of economizing but very much present in that of human action – the *entrepreneurial* element in human decision-making. It is this entrepreneurial element that is responsible for our understanding of human action as active, creative, and human rather than as passive, automatic, and mechanical. (Kirzner 1973, 35)

Even more broadly, Kirzner has defined entrepreneurship as the apprehension or identification of a means-ends framework, as distinct from maximization within a given configuration of means and ends (Kirzner 1979, 158).

Thus, the individual acts entrepreneurially when he finds new means of satisfying his wants, perhaps by discovering a previously unknown good, perhaps by finding a way to combine goods in a novel act of household production. Both of these activities involve finding new means of satisfying the individual's given ends, and thus they may not appear to constitute the "discovery" or "creation" of preferences. But we must recognize that the goods and services available in the marketplace are not, in and of themselves, want satisfaction; they are only means to want satisfaction. If we define an individual's preferences over goods and services, as is typical in microeconomic theory, we have to some extent assumed the individual has already discovered which goods and combinations thereof constitute the best means for the satisfaction of his more basic wants. If the process of learning about means occurs as the individual makes choices and interacts with the world, then preferences expressed as rankings of bundles of good will inevitably change over time, even if underlying wants remained fixed.

That preferences actually reflect a relationship between means and one's more basic ends has also been recognized by Austrian economists, as well as more mainstream

authors. O'Driscoll and Rizzo (1996) find an affinity between their approach to preferences and Becker's:

> …[W]ants are not directly for observable goods. Instead, they are for the satisfaction of some more basic desire: music (Stigler and Becker, 1977, p. 78), comfortable indoor temperature, health, and delicious meals (Becker, 1971, pp. 47-8) are all examples of basic human wants. (O'Driscoll and Rizzo 1996, 45)

As a result of discovery and error correction, the perceived connections between one's ends and the means of achieving them will necessarily change:

> As we have just seen, the relationship between wants and choices of market commodities is not a simple one. "There is many a slip betwixt cup and lip." The individual may incorrectly attribute want-satisfying potential to a good (Menger, 1981, p. 53; "imaginary good"), or he may ignore such potential when it is actually there. Furthermore, he may find that the costs of engaging in a particular plan of household production (in addition to the direct commodity costs) make the achievement of a particular want satisfaction prohibitively expensive. (O'Driscoll and Rizzo 1996, 45-46)

If this analysis is correct, then preferences over goods and services share some of the essential features of the plans formed by entrepreneurs and businesspeople. They are subject to error, correction, and modification over time as the agent learns more about both himself and his environment.

So far, I have emphasized the protean relationship between means and given ends; yet the ends themselves may also be subject to an ongoing discovery process. Even supposing the set of basic human wants (music, health, tasty meals, etc.) is both known and fixed, the relationship among them is not. The individual may not know, initially, his marginal rate of substitution between one goal and another. He may come to know it partially through an internal process of pure deliberation, but he may also experiment with different plans and activities, essentially "trying on" different sets of preferences over goods. This point, again, is anticipated by O'Driscoll and Rizzo:

> The individual, in his imagination, projects the likely consequences of different courses of action, including what must be sacrificed to achieve them. … In this process, the individual *clarifies his ranking of the imagined consequences*, his knowledge of the relationship between particular courses of action and those ranked consequences, and his perception of prices and income. (O'Driscoll and Rizzo 1996, 28; emphasis added).

Through a process of trial and error, the individual will tend to discovery strategies that suit him better.

To complicate matters, the very act of experimentation can sometimes change the underlying wants that are being discovered and clarified. In this sense, the individual not only discovers his wants and their relative importance, but also *creates* them. Contrary to the meta-preferences approach, however, no appeal to preferences over preferences is required to explain the fact that individuals make preference-affecting choices. Instead, preference-affecting choices are explained by the simple fact that preferences are never fully defined and known in advance, and therefore individuals must engage in a process of preference discovery whose unavoidable consequence is endogenous change in the preferences themselves. To the extent the agent is aware that his preferences will be affected by his actions, he may choose to take actions that will tend to reinforce or undermine some of his wants or desires. But the motivation for doing so derives entirely from the individual's existing wants and desires, as shaped by prior learning and experimentation.

Conclusions

Both meta-preference and multiple-selves approaches share a common error: identifying preferences and/or selfhood with a particular aspect of the individual. A

superior approach identifies preferences and selfhood with the process of interaction among different aspects of an individual.

Many of the phenomena that meta-preferences and multiple selves are meant to explain can be explained much more easily, and plausibly, in terms familiar to mainstream microeconomic theory. That individuals experience conflicting desires, that they experience regret over those desires left unfulfilled, that they resent the sacrifices required by the presence of constraints – these are all truisms, and they pose no problem for the traditional approach. Other phenomena – such as self-binding behavior and time inconsistency – are more difficult to reconcile with mainstream theory. But they, too, can be dealt with through specific exceptions to the usual microeconomic approach, specifically, a weakening (not abandonment) of the strong rationality assumption along the lines suggested by Elster and Schelling.

Meta-preferences and multiple selves have the appeal of familiarity. Humans beings *do* reflect on their preferences, and they *do* sometimes feel like different people argue within them. Yet these are only metaphors and personifications. The simplest way to see this is so is to ask *why* one has a particular meta-preference, or *where* a particular self-within-the-self came from. The answer, I suggest, is most often just a standard desire of the first order that is being frustrated in some way.

If we identify preferences and selfhood with a process rather than an outcome, the resulting model of human choice gains greater depth than the traditional microeconomic approach, without adding the unnecessary structure of multiple selves or multiple layers of preference. In this model, individuals seek to discover and refine their own preferences over time, through a process of experimentation and error correction. This

model undoubtedly has policy implications for such issues as advertising regulation, forced savings, availability of self-constraining devices, and so on. These issues largely overlap those that meta-preference and multiple-selves analysts have sought to address. Exploring the policy implications fully is beyond the scope of this paper. Here, I will only suggest that that while a process of discovery is never complete or error-free, systematic interference with that process could obstruct the experimentation and error correction that lead to improvement over time.

Figure 1: Meta-Preferences

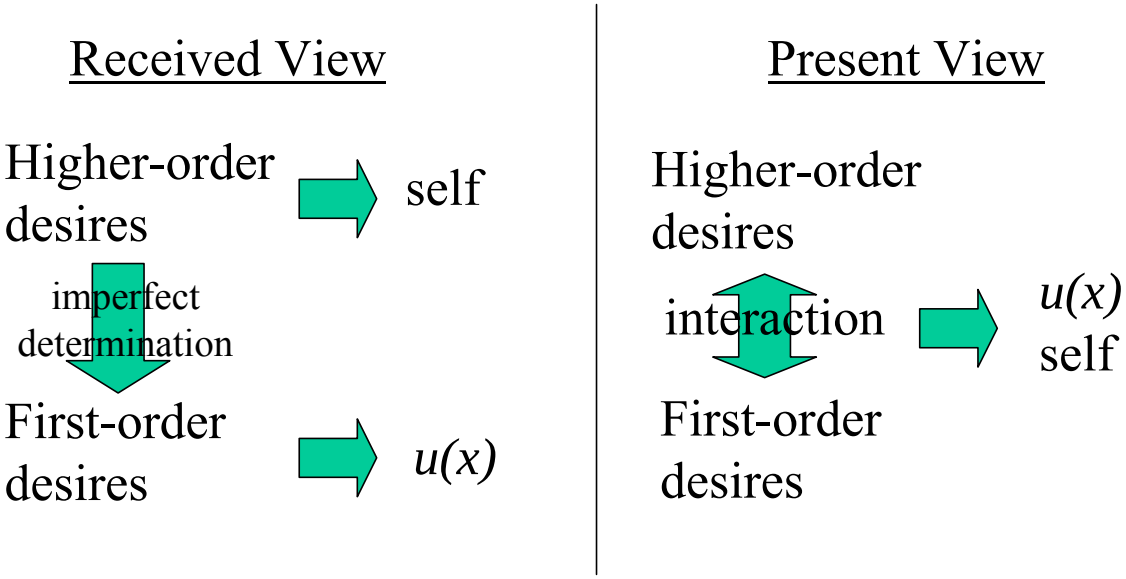
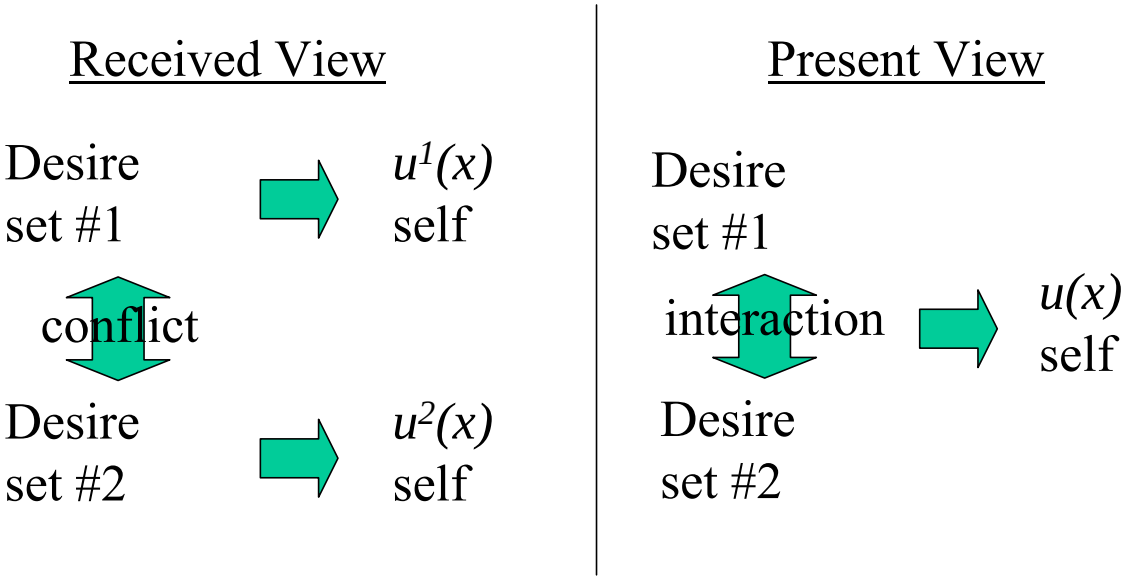


Figure 2: Multiple Selves



References

Becker, Gary S. (1965) "A Theory of the Allocation of Time," 75 Economic Journal 493-517.

Becker, Gary S. (1971) *Economic Theory*. New York: Alfred A. Knopf.

Becker, Gary S., and Murphy, Kevin M. (1988) "A Theory of Rational Addiction," 96 Journal of Political Economy 675-700.

Brennan, Timothy J. (1989) "A Methodological Assessment of Multiple Utility Frameworks," 5 Economics and Philosophy 189-208.

Buchanan, James M. (1979) *What Should Economists Do?* Indianapolis: Liberty Press.

Christman, John (1987) "Autonomy: A Defense of the Split-Level Self," 25 Southern Journal of Philosophy 281-293.

Christman, John. (1991) "Autonomy and Personal History," 21 Canadian Journal of Philosophy 1-24.

Cowen, Tyler. (1991) "Self-Constraint and Self-Liberation," 101 Ethics 360-373.

Cowen, Tyler (1993) "The Scope and Limits of Preference Sovereignty," 9 Economics and Philosophy 253-269.

Elster, Jon. (1979) *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Elster, Jon. (1985) "Weakness of the Will and the Free-Rider Problem," 1 Economics and Philosophy 231-265.

Frankfurt, Harry G. (1971) "Freedom of the Will and the Concept of a Person," 68 Journal of Philosophy 5-20.

Friedman, Marilyn A. (1986) "Autonomy and the Split-Level Self," 24 Southern Journal of Philosophy 19-35.

George, David. (1993) "Does the Market Create Preferred Preferences?" 51 Review of Social Economy 323-346.

George, David. (1998) "Coping Rationally with Unpreferred Preferences," 24 Eastern Economic Journal 181-194.

George, David. (2001) "Unpreferred Preferences: Unavoidable or a Failure of the Market?" 27 Eastern Economic Journal 463-479.

George, David. (2001) *Preference Pollution: How Markets Create the Preferences We Dislike*, Ann Arbor: University of Michigan Press.

Gifford, Adam. (2002) "Emotion and Self-Control," 49 Journal of Economic Behavior and Organization 113-130.

Gruber, Jonathan, and Koszegi, Botond (2000) "Is Addiction 'Rational'? Theory and Evidence," NBER Working Paper No. w7507.

Hayek, F. A. (1967) *Studies in Philosophy, Politics and Economics*. Chicago: University of Chicago Press.

Jeffrey, Richard C. (1974) "Preference among Preferences," 71 Journal of Philosophy 377-391.

Kirzner, Israel M. (1973) *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, Israel M. (1979) *Perception, Opportunity, and Profit*. Chicago: University of Chicago Press.

Kirzner, Israel M. (1985) *Discovery and the Capitalist Process*. Chicago: University of Chicago Press.

Klein, Daniel B. (1992) "Go Ahead and Let Him Try: A Plea for Egonomic Laissez-Faire," 35 Journal of Philosophy 3-20.

Laibson, D. I. (1997) "Golden Eggs and Hyperbolic Discounting," 112 Quarterly Journal of Economics 443-447.

Lancaster, K. J. (1966) "A new Approach to Consumer Theory," 74 Journal of Political Economy 132-157.

Lutz, Mark A. (1993) "The Utility of Multiple Utility: A Comment on Brennan," 9 Economics and Philosophy 145-154.

Mele, Alfred. (1993) "History and Personal Autonomy," 23 Canadian Journal of Philosophy 271-280.

Menger, Carl. (1981 [1871]) *Principles of Economics*. Ed. J. Dingwall; trans. B. F. Hoselitz. New York: New York University Press.

Mill, John Stuart. (1952) *Utilitarianism*. Great Books of the Western World, Vol. 43. Chicago: Encyclopaedia Britannica, Inc.

Mulligan, Casey S. (1996) "A Logical Economist's Argument Against Hyperbolic Discounting." Unpublished manuscript, University of Chicago.

O'Driscoll, Gerald P., Jr., and Rizzo, Mario J. (1996 [1985]). *The Economics of Time and Ignorance*. London: Routledge.

Rubenfeld, Jed. (2001) Freedom and Time: A Theory of Constitutional Self-Government. New Haven: Yale University Press.

Schelling, T. C. (1978) "Egonomics, or the Art of Self-Management," 68 American Economic Review, Papers & Proceedings of the 90th Annual Meeting of the American Economic Association, 290-294.

Schelling, T. C. (1996) "Coping Rationally with Lapses from Rationality," 22 Eastern Economic Journal 251-269.

Schutz, Alfred. (1967) *The Phenomenology of the Social World*. Trans. George Walsh and Frederick Lehnert. Evanston, Ill.: Northwestern University Press.

Stigler, George J., and Becker, Gary S. (1977) "De Gustibus Non Est Disputandum," 67 American Economic Review 76-90.

Strotz, R. H. (1955/6) "Myopia and Inconsistency in Dynamic Utility Maximization," 23 Review of Economic Studies 165-180.

Taylor, James Stacey (2003) "Autonomy, Duress, and Coercion," 20 Social Philosophy and Policy 127-155.

Thaler, Richard H., and Shefrin, H. M. (1981) "An Economic Theory of Self-Control," 89 Journal of Political Economy 392-406.

Weisbrod, Burton A. (1977) "Comparing Utility Functions in Efficiency Terms or, What Kind of Utility Functions Do We Want?" 67 American Economic Review 991-995.