# COPING RATIONALLY WITH UNPREFERRED PREFERENCES

**David George**
*La Salle University*

The fact that people apparently can, by their own account, behave "irrationally" has been a point of tension for economists. Responses have, not surprisingly, differed. Following a variation of Alfred Marshall's vision of economics, some have chosen to fully accept "irrationality" as a reality but to leave it outside the scope of economics. By this approach, the "economy" is precisely that sphere in which forethought and careful choice do in fact prevail. As a different strategy, others have chosen to simply tautologize the rationality claim. One, by definition, does what one prefers (from among the options available), and cannot possibly be in any sense irrational.[1]

In more recent years, two more developed sorts of responses to apparent irrationality have emerged. One approach, of which Thomas Schelling's work provides one example, is that of "dual" or "multiple" selves, within the single person.[2] Paradoxical instances of an agent restricting her choice set have been conceptualized as attempts by one such "self" to impose its preferences on a contending internal self. The other approach has been that of "metapreferences" or "second-order preferences."[3] Within this tradition, an agent has at any moment not only a preference ranking of the opportunities available but a ranking of different potential rankings as well.

My attempt in this paper will be to strengthen the case for the second-order preference approach vis-à-vis the multiple-selves approach. My plan is as follows. In the opening section I summarize two particular sorts of internal conflict recently presented by Schelling [1996] in this *Journal*. While the normative coherence of his approach works well in the first of the cases summarized, I show that it falls decidedly short in the second. In the section that follows, I describe second-order preferences and emphasize their superiority to the Schelling approach. Section three will review some of the economics literature that has placed unnecessary limiting conditions on what it means to say that an agent's preferences have improved. The section that follows demonstrates how the second-order preference approach provides the means of evaluating preference change that is lacking in the Schelling approach. In the final section, second-order preferences are shown to represent something of a compromise between the position represented by Aristotle that preferences can be objectively ranked and the position represented by most contemporary economists and popular social thought that they are not even subjectively rankable.

## SCHELLING'S TYPES OF INTERNAL CONFLICT

Schelling's interest in internal conflict has a long history.[4] The remarks to follow are based on his recent article in this *Journal*, a wide-ranging paper that seeks to distinguish between clearly different sorts of conflict within the person. The most successful use of Schelling's analytical structure occurs when the agent is choosing to protect himself from a future "unconscious self" or clearly incompetent self (e.g., one's future intoxicated self) by imposing constraints. A far less successful use of this structure occurs when the potentially wayward future self whom the present self is attempting to influence would be fully conscious and competent at the time of carrying out the act against which the present self is seeking to protect her. I will now offer a closer look at each of these cases.

*CASE 1.* The example that Schelling offers conforms closely to the experience of many who are trying to give up smoking. The events that are likely to precipitate the urge to smoke are many, and drinking stands as one of the most often mentiond activities that serves to undermine the quitter's efforts. In Schelling's words: "One possibility is that he anticipates drinking alcohol and knows by experience that drinking interferes with his reasoning so that he succumbs to the temptation of cigarettes if cigarettes are readily available. If the individual believes that what alcohol does is to distort or to anesthetize some of the functions or characteristics that go under the name of rationality, we can call this situation 'anticipated irrationality'"[1996, 265].

Formally speaking, acceptance of preemptive strikes on the part of the agent to prevent future smoking that he would voluntarily undertake if free to do so requires only a relaxing of the assumption that preference is always revealed when choice is exercised. That is, allowing N to represent "not smoke" and S to represent "smoke", we have an agent whose preference is "N pref S" but who acts irrationally when intoxicated. To rule out the very possibility that S will be chosen, it is removed from the choice set, and the agent in a rather indirect way "chooses" the preferred action, N, in this manner. This is summarized as Case 1 in Figure 1.

It is readily apparent that Schelling's analytical structure serves quite well in this first case. Constraining future choices is a means of choosing that which is now and will later be the "rational thing to do," but which the agent will later simply be unable to do.

*CASE 2.* Schelling considers separately a case in which the agent's preference and not just his choice depends on the time separating choice from action. When sober, the agent prefers to not smoke and thus refrains. When intoxicated, it is the reverse. By this account, "the craving is such a discomfort and distraction that it will produce an unproductive and disagreeable evening; at the risk of complete relapse it would be rational to avoid such a dismal and unproductive evening by going ahead and smoking" [1996, 266]. It is critical to recognize that in Case 1 the agent's ranking of possible choices did not change as a consequence of his preemptive strike to limit future choices. His motive for making the decision well prior to the actual time of action was that he would not act rationally at this later time. Here, in contrast, we have the agent ruling out as a possibility the choice that he would prefer were he not to take any action. This Case 2 is also summarized in Figure 1.

The question marks that appear in the final row reveal a shortcoming of the analytical structure that Schelling chooses to offer. Superficially, it may appear that a welfare judgment is indeed possible. For it is described as a "craving" to smoke that the agent experiences when cigarettes are available. To emphasize that cravings seem to delegitimize preferences I have offered a variation of Case 2 in Figure 1. Let "smoking and accepting a punch" be a third alternative and let it be ranked in the second position at the potential time of smoking on Saturday, ahead of "not smoking," while being ranked dead last on Thursday.

**FIGURE 1**

| | Case 1 | Case 2 | Variation on Case 2 |
|---|---|---|---|
| **1. Preference for time 2 at time 2** | N pref S | S pref N | S pref (S + punch) pref N |
| **2. Choice for time 2 at time 2** | S | S | S |
| **3. Explanation for choice** | "irrational" | preferred preference not an option | rational choice |
| **4. Preference for time 2 at time 1** | N pref S | N pref S | N pref S pref (S + punch) |
| **5. Choice for time 2 at time 1** | N | N | N |
| **6. Welfare comparison of "5" with "2"** | 5 superior | ?? | ?? |

At first consideration this throws open to doubt the legitimacy of the preference for smoking. On further consideration, however, it is apparent that this "craving" is not a sufficient criterion for delegitimizing a preference. To see this, let S represent "engage in creative writing" and N represent the opposite ("not engage"). One might look with admiration and respect on someone summarized by lines 1 and 2 of "the variation of Case 2." For here would be an agent having such a strong preference to write that not even a pending punch would cause her to give up her efforts. Clearly, "cravings" are not necessarily states that we evaluate as unfortunate. While Schelling seeks to rationalize the efforts of an agent to overcome such a craving, his example has appeal only on an *ad hoc* basis. At a more abstract level, unlike Case 1, the agent's action cannot be judged to be necessarily welfare enhancing.

## SECOND-ORDER PREFERENCES

One way out of this normative impasse is to allow for the possibility that one who is restricting one's future freedom is motivated by a second-order preference. While a first-order preference is over "bundles" or "activities," a second-order preference is necessarily over "rankings of bundles" or "rankings of activities." To have a second-order preference to not smoke does carry with it the implication that the agent is better off having a first-order preference to not smoke. But to tell whether the agent is better off smoking or not smoking, we need to know what the agent's *first*-order preference actually is. By definition, if it is to smoke, she is better off smoking than not smoking. And if the restriction of future choices manages to shape the first-order preference as she wishes to have it shaped, such self-imposed restrictions make sense.

That humans have the potential to have preferences about preferences is not a new idea, as a consideration of Aristotle in the paper's final section will be intended to demonstrate, but acceptance of such a human proclivity among economists has been tangential at best. In a seminal work by the philosopher Harry Frankfurt, humans were not only argued to have the *ability* to have preferences for preferences but this attribute was argued to be the defining feature of personhood [1971, 6]. Animals indeed prefer this to that but animals, argued Frankfurt, are without the ability to step back and judge their preferences. While the arguments among philosophers about the full significance of second-order preferences have been considerable, none, to my knowledge, has denied Frankfurt's essential claim that only humans can indeed reflect on their tastes.

A functioning economy, as usually defined, is a uniquely human institution, and it is noteworthy that this is so within the standard economic model only because humans have developed the institution of private property and have exhibited the ability to specialize and trade. On the consumer front, the implied posture of *homo economicus* toward his desires is the same as an animal's posture toward its desires. More to the point, neither is assumed to have the capacity to evaluate its preferences, and to embrace or seek to change these preferences following such an evaluation. Some introductory texts have sought to convey the full generality of demand theory by reporting empirical findings that animals have downward sloping demand curves as surely as do humans [Ruffin and Gregory, 1988, 4; and Boyes and Melvin, 1991, 150]. So, for example, if the number of pecks of a bar that are required to issue forth a fixed quantity of food is increased, then the representative pigeon will, just as positive theory predicts, opt for less. The rise in the "price" will have lowered the "quantity demanded." If Frankfurt is correct, this has to give economists pause, at least in their normative capacity. While change in the well-being of the consumer serves as the ultimate measure of economic improvement, this consumer is no different than what Frankfurt chooses to call a "wanton," one who, like lesser creatures, doesn't reflect on his tastes.[5]

Frankfurt suggests we might look at this in another way. It is common to refer to a set of potential choices facing an agent and to say that she has a "free will" if she is free to choose that which she most prefers. So, for example, if a struggling economics student has the option of (1) studying for a pending economics test, or (2) visiting some friends, her decision to select the second of these is an instance of the exercise of a free will. For Frankfurt, a choice of an activity (or bundle of goods) is better understood as the "freedom to act upon the will that one has." Whether the will is free or not hinges critically upon whether or not the agent would have chosen the preference ranking that she in fact happens to be experiencing. If this struggling student wishes that she had more of a taste for economics, Frankfurt would say that her will was not free, but would only be so if she preferred studying over visiting friends.

Now while it should be fairly apparent that I believe economists would benefit from paying heed to what someone outside their ranks is saying, it is only fair to note that Frankfurt's argument might have been more compelling had it been modified with some fairly standard economic insights. To speak of having the "freedom to act upon one's will" and to speak of having a "free will" is to be too categorical. Starting with the first, the wayward student might have preferred partying to either of the already mentioned options, but this might not have been an option on this particular occasion. It follows that she has a circumscribed "freedom to act on the will that she has," meaning simply that a limited number of options are available. We can speak of

more or less "freedom to act" as the choice set widens or narrows. By the same reasoning exactly, "free will" becomes a bit strong. To have the preference ranking that one ranks as the very best imaginable would be superhuman, but to have one that is better than what one currently has would not be. Hence, for this agent to be able to change her preference ranking such that studying is preferred to the other alternatives might be better described as a case of this agent having a "freer will" than with what she is currently saddled.

## AN OVERLY RESTRICTIVE INTERPRETATION

In articles appearing in the 1970s, something much like a second-order preference ranking figured in the normative economics literature on two different fronts. Burton Weisbrod [1977] worked within the more traditional format with a utility function rather than a preference ranking as his engine of analysis. Amartya Sen [1974; 1977], in contrast, relied on preference rankings rather than the utility function, and at least partly as a result of this managed to draw philosophers and decision theorists into the discussion.

The title of Weisbrod's article, "Comparing utility function in efficiency terms or, what kind of utility functions do we want?" creates an initial impression that he is receptive to second-order preferences. His loyalty to orthodoxy does, however, lead him to greatly qualify the extent to which it is possible to rank utility functions. For he concludes that "The customary proposition that one type of utility function cannot be compared to another within an economic efficiency framework is correct in general" [1977, 994]. He further concludes that some very stringent conditions must be met before a comparison of utility functions is even possible, namely, that ". . .one type of utility function, and the expected consumption bundle it generates, may be said to be preferred to another, and the expected consumption bundle it generates, if and only if (a) the two expected consumption bundles are different and (b) the same consumption bundle is preferred no matter which utility function is used to evaluate the two bundles" [1977, 993].

It is the second of these conditions that is not at all supported by the model that I am presenting and on which attention will be focused for the remainder of this section. In order to give an illustration of what the fulfillment of this condition entails, it is necessary to expand the two-element ("smoke" or "don't smoke") example that has been used to this point in the paper. Let there now be three elements, A, B and C. Suppose that when the agent has (A pref B pref C) as her preference, this somehow *causes* elements A and B to not be available and that as a consequence, C is her choice. Further suppose that when her preference shifts to (B pref A pref C), B *is* suddenly available and becomes her choice. According to Weisbrod, the latter preference ranking would be preferred by the agent since the choice that follows from it is preferred to the choice accompanying the other preference, and this preference holds *whichever of the two rankings one refers to.* The agent's second-order preference for (B pref A pref C) over (A pref B pref C) is thus strictly instrumental in nature. It amounts to a strategy for being able to attain a more highly-ranked element in the original preference ranking, and is not in any sense an exercise in embracing or rejecting this original ranking qua ranking.

It should be apparent that this condition is not fulfilled by the second-order preference approach as developed thus far. To say that a smoker has a second-order preference to not smoke has been taken to indicate that the person is better off "not smoking while preferring to not smoke" than he is "smoking while preferring to smoke." It is most certainly not the case that the same act was preferred both before and after the preference change but for some reason not attained prior to the change. Weisbrod's criterion contradicts a clear implication of the second-order preference that has been presented here. Despite this, similar claims appear elsewhere in the literature [Frank, 1987; Dixit and Norman, 1978, 2; and Cowen, 1993, 256].

Operating within the structure of preference rankings rather than utility functions, Amartya Sen places no similar requirement on second-order preferences. In an article appearing several years before Weisbrod's, Sen presents a guilt-ridden meat-eater saying "'I wish I had a vegetarian's tastes, for I disapprove of the killing of animals, but I find vegetarian food so revolting that I can't bear to eat it, so I do eat meat'" [1974, 63]. Letting M stand for "eating meat" and V stand for "eating vegetables," this agent is reporting that (1) [(V pref M) pref (M pref V)] is his second-order preference, (2) that if the (V pref M) were his first-order preference, V would be his choice, (3) but since (M pref V) happens to be his first-order preference, M is what he chooses. Within much of Sen's writings on second-order preferences, to simply assert that the agent has a second-order preference for one thing over another carries with it an implicit assumption that the elements within the first-order ranking are within the agent's choice set, both before and after any preference change. Quite clearly, he does not adhere to the Weisbrod criterion, for it is not required by him that "the same consumption bundle is preferred no matter which utility function is used to evaluate the two bundles."

In spite of this clear difference, the context in which Sen first introduced the second-order preference happens to fulfill the Weisbrod conditions. Behaving "as if" one had a preference other than what one actually had was argued by Sen to be a device for overcoming the Prisoner's Dilemma. To summarize his argument, Figure 2 presents the fabled dilemma. The number on the left in each of the boxes represents agent 1's payoff, the number on the right agent 2's payoff. If agent 1 is rational and self-interested, his first-order preference ranking (stated in terms of his payoffs) would naturally be (4 pref 3 pref 2 pref 1) which would lead to the selection of clearly dominant strategy B. If agent 2 had an equivalent ranking of her potential payoffs, she too would select B and the well-known result would be a payoff of 2 for each agent. As Sen argues, if each agent were to behave "as if" he or she valued the other's well-being in addition to his or her own, the dilemma might be overcome. To take just one of the examples considered by Sen [1974, 60-62], each agent might prefer having "other-regarding preferences" that lead to the ranking of outcomes on the basis of the sum of payoffs enjoyed. This would result in first-order preference ranking (3 pref 1 pref 4 pref 2) for each agent (the agent's personal payoff alone is still being used to express each outcome), the selection of A by each (the now dominant choice) and the attainment of a payoff of 3 for each. For each agent, the first-order preference and outcome

would have changed as shown in Figure 3. What is significant to note is that 3 ranks higher than 2 for both the "selfish" and the "other-regarding" agent. Since a different "bundle" (in this case, "payoff") is realized following the changed ordering, it follows that both of the Weisbrod criteria are fulfilled.

FIGURE 2

| | | AGENT 2 | |
|---|---|---|---|
| | | A | B |
| AGENT 1 | A | 3,3 | 1,4 |
| | B | 4,1 | 2,2 |

FIGURE 3

| | All Agents Selfish | All Agents "Other-Regarding" |
|---|---|---|
| **1. Preference for each** | 4 pref 3 pref 2 pref 1 | 3 pref 1 pref 4 pref 2 |
| **2. Choice for each** | B | A |
| **3. Payoff for each** | 2 | 3 |

Sen's decision to introduce second-order preferences was, it should be emphasized, *ad hoc* in nature. It was not from a pure methodological standpoint that he proceeded but from the desire to give a richer rendering of the way in which Prisoner's Dilemmas could be overcome. While it is clear that he did not, like Weisbrod, treat the second-order preference as coherent only if it allowed a higher element within the unchanging first-order preference ranking to be attained, it may have helped to strengthen this misinterpretation.

## REINTERPRETING SCHELLING'S CASE 2

In Figure 4 appears another way of showing Schelling's conflicted person who was summarized by Case 2 in Figure 1. Although the second-order preference, first-order preference, and choice are separated by three days, they pertain to the same time, namely, let us say, Saturday evening. Note first that the agent's second-order preference is assumed not to change. Whether it is Wednesday or Saturday, he would prefer to have a preference on Saturday to abstain from smoking. His first-order preference and choice of action do, however, differ. On Wednesday he prefers to not smoke on Saturday and would thus choose, if it were possible, to rule out smoking on Saturday. On Saturday, in contrast, his preference would be to smoke and that is what he would choose to do if the opportunity presented itself.

FIGURE 4

| | Wednesday regarding Saturday | Saturday regarding Saturday |
|---|---|---|
| 1. Second-order | (N pref S) pref (S pref N) | (N pref S) pref (S pref N) |
| 2. First-order | N pref S | S pref N |
| 3. Choice | N | S |

As some introspection will reveal, the pattern shown in Figure 4 is not an unusual one. It is much more common for the first-order preference to shift as the time between decision and consumption narrows than it is for the second-order preference to shift. When trying to diet I begin the day preferring to prefer a light dinner and that second-order preference doesn't budge. It's a different story with the first-order preference. Early on my preference is to indeed have a light dinner, but as evening approaches, the preference to eat quite a bit more arises and, saddled with it, I act upon it.

Even granting this general pattern for those areas of consumption that might be called addictive or compulsive, is it possible to say that an agent who is able to act early rather than late is making himself better off? Schelling's approach certainly rules this conclusion out, and mine appears to be no better as it presently stands. If the agent is able to act on Wednesday, she will on Saturday have a second-order preference to *not* smoke, and a first-order preference *to* smoke, but would have ruled out smoking as a possibility. This would have to leave her worse off. This is a claim that may run counter to one's intuition for at least two reasons.

First, a second-order preference is sometimes interpreted to be nothing more than a "superior" ranking of the same elements that appear in the first-order preference. For one so reasoning, an agent who rules out smoking manages to get the preference of her "higher" self (N pref S) to take precedence over the preference of her "lower" self, (S pref N). But this is a mistake. As noted earlier, it only makes sense to choose that activity which one "prefers to prefer" if one fortunately happens to prefer this activity.

A second point of confusion has nothing to do with second-order preferences directly, but with the assumption that an agent, when thoughtful and deliberate and not in a state in any way resembling Schelling's Case 1, always chooses to do what he prefers. If by preference we intend the "overall" sense of the word, then the agent is indeed best off doing what he prefers. If we are using the word more narrowly, this is not necessarily so. To see the difference, consider someone reporting that although he prefers the cholesterol- and calorie-riddled cheese lasagna to anything else on the menu he is going with a healthier choice. In this context, (a perfectly fine use of "prefer" in everyday life), the context is narrowed considerably. A person reporting these food tastes can be interpreted to be indicating that the next 30 minutes will be most enjoyable if he goes with the lasagna since the benefits are up front and potential costs far removed. But in an overall sense, his preference is for the healthier choice, not for the lasagna. And it is the overall sense that has always prevailed (implicitly) in choice theory. Thus, the agent having a preference to smoke is, *by definition,* better off smoking than not, regardless of whether or not his second-order preference is to have a different preference.

FIGURE 5

| | Second-Order Preference | First-Order Preference | Choice |
|---|---|---|---|
| A | (N pref S) pref (S pref N) | S pref N | S |
| B | (N pref S) pref (S pref N) | S pref N | N |
| C | (N pref S) pref (S pref N) | N pref S | S |

This now returns us to the original question. Does the agent's pre-emptive strike of ruling out cigarettes leave him better off or not? The answer, at least tentatively, would appear to hinge on whether or not the agent succeeds in changing his preference on Saturday as a consequence of this self-imposed shrinking of the choice set. In Figure 5 appear three combinations of second-order preference, first-order preference and "choice." The first combination, A, shows the situation if the agent had taken no action on Wednesday. Though preferring to have a preference to abstain from smoking, his first-order preference is to smoke and he consequently does smoke. Combination B summarizes the situation when the Wednesday action rules out cigarettes, but has no effect on his first-order preference on Saturday. Finally, combination C shows what would have to be judged a successful outcome. The action on Wednesday has managed to not only keep the agent from smoking on Saturday but has caused his tastes to turn against smoking. Through what Jon Elster has usefully called a "sour grapes" reaction, the agent, unable by his own Wednesday decision to have cigarettes on Saturday, has found himself really not missing them at all. By ruling out the possibility of having them, he has shaped his first-order preference against them.

An important conclusion follows: within 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. Clearly, state of affairs C is preferable to A which is in turn preferable to B. In other words, having the preference one prefers to have and acting on it is preferable to having the preference one doesn't prefer and acting on it. The latter, in turn, is preferable to having the unpreferred preference be a reality but to be at the same time unable to act on it. By this analysis, the act of ruling out that which the agent would prefer not to prefer is anything but neutral. It will leave the agent better off or worse off than he would otherwise be, depending on the action's success in shifting the agent's first-order preferences.

And this raises an interesting distinction between Schelling's Case 2 and the alternative model I am proposing. By the Schelling account, whether or not the conflicted agent is able to successfully limit the *choice* set is the determining factor in whether or not he will intervene. By my account, whether or not he can change the *first-order preference* is the determining factor. Let me give a personal example that may illustrate the difference. I have some eating habits with which I am particularly unhappy. When I was single I never, ever, allowed myself or others to bring potato chips into my apartment. To do so was to assure their disappearance in a ridiculously short time. Despite a craving for chips when they were available, I don't recall missing them at all when they were not. The act of deciding on food shopping day to not purchase them succeeded not only in stopping me from eating them, but also shifted my preferences for the better.

Sleeping provides a very different case. While I certainly enjoy sleeping, at times I would prefer having a preference for less than my usual seven hours. While I prefer the seven I would prefer to prefer, say five hours. Despite this, never would I try to devise strategies for sleeping less through pre-commitment. To do so would not affect my preference for sleep in the slightest. Referring again to Figure 5, I would be in a case B situation, worse off than if I just went with the existential case A reality, recognizing that the human condition sometimes necessitates that we have preferences that are other than what we would wish.[6] And it is only a definitional problem that might appear to call this claim into question. Admittedly, many dieters and ex-smokers will maintain that they are doing what they are *despite* having a strong preference to eat or to smoke. But this "preference" would, on closer examination, be in the nature of an "intrisic" rather than an "overall" preference. Such an agent is best understood to be defying her intrinsic first-order preference as well she should since her overall first-order preference points in a different direction.

## ARISTOTLE'S CATEGORIES

The orthodox economic vision of rational action agrees in certain important respects with the vision of rationality offered by Socrates, for whom "bad" choice could only result from a less-than-full grasp of the facts. In the works of Aristotle, in contrast, the sorts of internal conflict being explored here were much in evidence. In Book VII of the *Nicomachean Ethics* [Ackrill, 1973], Aristotle distinguishes between four types of people, three of which will be considered here; the "temperate," the "incontinent," and the "self-indulgent." Letting I = act in accord with impulses, and E = exercise restraint, the combination of second-order preference, first-order preference, and action for these three types of agent are summarized in Figure 6.

While it would be possible to interpret Aristotle's distinctions in other ways, second-order preferences provide a particularly clear way of highlighting the differences. For "Temperate" the first-order preference is (E pref I) and the chosen action thus E. For both "Incontinent" and "Self-indulgent" the first-order preference is (I pref E) and the selected action is I. "Temperate" and "Incontinent" share a second-order preference to exercise restraint, while "Self-indulgent" stands alone in having a second-order preference to go with his impulses.

There may appear to be nothing particularly different than what has already been discussed in this paper. "Incontinent's" second-order preference, first-order preference and choice closely resembles what the frustrated smoker's profile would be on Saturday were he not to take pre-emptive action. "Temperate" and "Self-indulgent," in contrast, are each content with the preference that they have. But clearly something new has been added to the picture. The name "Temperate" accords well with an agent who is at peace with her preferences, while the name "Self-indulgent" opens up

**FIGURE 6**

| | Second-order Preference | First-order Preference | Choice |
|---|---|---|---|
| **Temperate** | (E pref I) pref (I pref E) | (E pref I) | E |
| **Incontinent** | (E pref I) pref (I pref E) | (I pref E) | I |
| **Self-Indulgent** | (I pref E) pref (E pref I) | (I pref E) | I |

**FIGURE 7**

| | Neoclassical | Second-order pref | Aristotle |
|---|---|---|---|
| **1) Can bundles be ranked?** | Yes | Yes | Yes |
| **2) Criterion for ranking** | Agent's tastes | Agent's tastes | Objective |
| **3) Can preferences be ranked?** | No | Yes | Yes |
| **4) Criterion for ranking** | — | Agent's tastes | Objective |
| **5) Can second-order preference be ranked?** | — | Yes | Yes |
| **6) Criterion for ranking** | — | Agent's tastes | Objective |

a whole new issue. It seems that Aristotle was evaluating an agent's *second*-order preference. While "Temperate's" second-order preference was judged to be admirable, "Self-indulgent's" second-order preference was not.

If this interpretation is correct, it would seem to follow that the second-order preference structure as here presented stands somewhere between Aristotle's belief that preferences can be evaluated and the orthodox economic belief that they cannot be in any way evaluated. This is summarized in Figure 7. In line two appears the first difference between the three. The second-order preference approach, at least as developed here, would have to be aligned with the neoclassical approach by this criterion since for each, the relative goodness of something is determined by the preference of the choosing agent. For Aristotle, in contrast, there is a criterion that transcends the individual's preferences. With line three, the second-order preference approach shifts its loyalties. Contrary to what the neoclassical approach allows, but in accord with what Aristotle allows, evaluation of preferences is possible. In line four, the second-order approach assumes an independence. It is unlike the neoclassical approach in offering an evaluation of the agent's preferences. In is unlike the Aristotelian approach in having the agent be the evaluator of her preferences rather than having them be capable of objective evaluation.

The entries that appear in lines 5 and 6 are a repeat of lines 3 and 4, and from this it follows that a re-thinking is necessary as to what distinguishes Aristotle's "self-indulgent" from the agent considered throughout the paper who happened to have the preference that she preferred having. For just as surely as the second-order approach recognizes the agent's capacity to evaluate her preferences, there is nothing stopping us, on a formal level, from permitting the agent to evaluate her second-order preference. While Aristotle might be said to have *required* an evaluation of the second-order preference, the second-order preference model itself can at least recognize that it on occasion occurs.

Frankfurt chose to emphasize that discord between second-order and third-order is a rarity.[7] It is worth emphasizing, in closing, that there are some instances when an agent experiencing discord between the first-order and second-order might wish to change the latter rather than the former. For example, an "in the closet" gay person who prefers men to women (and acts on this), and a socially inculcated second-order preference to prefer to prefer women might wish to see it be the second-order preference that is altered. This would suggest a third-order preference (he prefers that he would prefer to prefer men) and does, of course, complicate the story considerably. As I have argued elsewhere, following Frankfurt, such cases of conflicts at a higher-order are rare, particularly in the realm of decision making economists are prone to focus. Further consideration of such deeper conflicts can remain the sphere of depth psychologists, or perhaps await consideration by economists at a future time.

## CONCLUSION

The approach offered by Thomas Schelling for understanding self-imposed constraints is well-suited for those instances when the rationality assumption clearly breaks down. The intoxicated person acting impulsively and the sleeping person acting with no self-awareness have been rightly removed by Schelling from the realm of rational choice. It has been my main argument that Schelling has been far less successful in attempting to use the same analytical structure to also account for self-imposed constraints that are, as he himself acknowledges, considerably more paradoxical. An agent's choice to rule out now the possibility of eating high calorie foods in the future is not (at least in my own case) an example of the present self exercising control over a future self who will be clearly "out of control," but is instead an example of the agent's present-self exercising control over a future calm, deliberate consumer.

Second-order preferences have been offered as an explanatory device that better captures such internal conflict and that provides a context in which self-imposed constraints can be shown to have clear normative implications. Provided that a self-constraint succeeds in altering not just future consumption but the agent's future preference ranking as well, it has been shown that a clear welfare gain could result. No longer must the study of internal conflict be viewed as a redistributional affair, with one "self" gaining at the expense of another "self." A united self is again in the picture, and this united self can be understood as exercising welfare-enhancing strategies.

To evaluate preferences makes many economists uneasy, particularly those for whom the acceptance of preferences as exogenous forms the very core of the discipline. There is more than a little irony in such a position. At the very time that the rational choice model is extending well beyond the traditional realm, it would be odd indeed if a preference that has been isolated by one philosopher as the single characteristic most distinguishing humans from other living creatures—the preference for one's preferences—were declared to be outside the domain of economics.

## NOTES

The author wishes to thank Robert Goldfarb, Timur Kuran, and an anonymous referee for valuable comments on an earlier version of this paper. Thanks also to Marc Santugini for the technical assistance.

1. For a sophisticated defense of the position that agents, by definition, "maximize," see Boland [1981]. For a rigorous analysis of contending definitions of rationality, see Walsh [1996, chapters 2-5].
2. Harsanyi [1955] is perhaps the originator of this tradition. Other contributors include Schelling [1978; 1980; 1984a; 1984b; 1996], Elster [1979; 1982; 1985; 1989], Thaler and Shefrin [1981], Loewenstein and Thaler [1989], Cowen [1991; 1993], and Ainslie [1992]. See Brennan [1989; 1993] for a critical review of the "multiple-utility" frameworks and a defense of the orthodox "single-utility" assumption. For a critique of Brennan's argument, see Lutz [1993].
3. Frankfurt [1971] is generally credited with being the first to develop in any detail the notion of a second-order preference. For other works on the subject, see Sen [1974; 1977], Jeffrey [1974], George [1978; 1984; 1993], Majumdar [1980], McPherson [1980; 1982], Van der Veen [1981], Hahn [1982], Hirschman [1982, chapter 4; 1985], Etzioni [1986], and Lutz [1993].
4. See Note 2.
5. Amartya Sen [1977], in choosing to call the standard *homo economicus* a "rational fool," was approximating Frankfurt's description of the "wanton." But unlike Frankfurt, Sen's underdeveloped agent wasn't characterized by the failure to evaluate her tastes, but rather by a failure to settle on a most strategic way to satisfy them. More will be said on this later in the paper.
6. This raises another complication. If I know with certainty that I will have a preference for X at the time of actual consumption (e.g., a preference to sleep at 11:30 on Thursday) I should have this same preference for this time at any point prior, regardless of what my second-order preference is. Were it otherwise, we would be back to a "two-selves" model, and all the problems that can be thus created. In short, given that I have a second-order preference for Y but a first-order preference for X at the time of consumption, a first-order preference for Y at any time prior only makes sense if by acting on it I impose upon myself not just Y but a preference for Y.
7. After acknowledging that "[A] person may have. . .desires and volitions of a higher other than the second," Frankfurt specified that it was the more decisive second-order desires that he was interested in. As he put it, "When a person identifies himself *decisively* with one of his desires, this commitment 'resounds' throughout the potentially endless array of higher orders" [1971, 16].

## REFERENCES

**Ackrill, J. L., translator.** *Aristotle's Ethics.* New York: Humanities Press, 1973.

**Ainslie, G.** *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person.* Cambridge: Cambridge University Press, 1992.

**Boland, L. A.** On the Futility of Criticizing the Neoclassical Maximization Hypothesis. *American Economic Review,* December 1981, 1031-36.

**Boyes, W. and Melvin, M.** *Microeconomics.* Boston: Houghton-Mifflin, 1991.

**Brennan, T. J.** A Methodological Assessment of Multiple Utility Frameworks. *Economics and Philosophy,* October 1989, 189-208.

__________. The Futility of Multiple Utility. *Economics and Philosophy,* April 1993, 155-64.

**Cowen, T.** Self-Constraint versus Self-Liberation. *Ethics,* January 1991, 360-73.

__________. The Scope and Limit of Preference Sovereignty. *Economics and Philosophy,* October 1993, 1-18.

**Dixit, A. and Norman, V.** Advertising and Welfare. *Bell Journal of Economics,* Spring 1978, 1-18.

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

__________. Sour Grapes - Utilitarianism and the Genesis of Wants, in *Utilitarianism and Beyond,* edited by A. Sen and B. Williams. Cambridge University Press, 1982, 219-38.

__________. Weakness of Will and the Free-Rider Problem. *Economics and Philosophy,* October 1985, 231-65.

__________. *The Cement of Society.* Cambridge: Cambridge University Press, 1989.

**Etzioni, A.** The Case for a Multiple-Utility Conception. *Economics and Philosophy*, October 1986, 159-83.

**Frank, R. H.** If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience? *American Economic Review,* September 1987, 593-604.

**Frankfurt, H. G.** Freedom of the Will and the Concept of a Person. *The Journal of Philosophy,* January 1971, 5-20.

**George, D.** The Market System and Second-Order Wants. *The Forum for Social Economics,* Spring 1978, 42-44.

__________. Metapreferences: Reconsidering Contemporary Notions of Free Choice. *International Journal of Social Economics,* 1984, 92-107.

__________. Does the Market Create Preferred Preferences? *Review of Social Economy,* Fall 1993, 323-46.

**Hahn, F.** On some Difficulties of the Utilitarian Economist, in *Utilitarianism and Beyond,* edited by A. Sen and B. Williams. Cambridge, Cambridge University Press, 1982, 187-98.

**Harsanyi, J. C.** Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility. *Journal of Political Economy,* August 1955, 309-21.

**Hirschman, A. O.** *Shifting Involvements: Private Interest and Public Action.* Princeton: Princeton University Press, 1982.

__________. Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse. *Economics and Philosophy,* April 1985, 7-21.

**Jeffrey, R. C.** Preferences among Preferences. *The Journal of Philosophy,* July 1974, 377-91.

**Loewenstein, G. and Thaler, R. H.** Intertemporal Choice. *Journal of Economic Perspectives,* Fall 1989, 181-93.

**Lutz, M. A.** The Utility of Multiple Utility: A Comment on Brennan. *Economics and Philosophy,* April 1993, 145-54.

**Majumdar, T.** The Rationality of Changing Choice. *Analyse & Kritik,* 1980, 172-78.

**McPherson, M. S.** Liberty and the Higher Pleasures: in Defense of Mill. *Economic Inquiry,* April 1980, 314-20.

__________. Mill's Moral Theory and the Problem of Preference Change. *Ethics,* January 1982, 252-73.

**Ruffin, R. J. And Gregory, P. R.** *Principles of Economics,* 3rd ed. Glenview, IL: Scott, Foresman, and Company, 1988.

**Schelling, T. C.** Egonomics, or the Art of Self-Management. *American Economic Review: Papers and Proceedings,* May 1978, 290-94.

__________. The Intimate Contest for Self-Command. *The Public Interest,* Summer 1980, 94-118.

__________. *Choice and Consequence.* Cambridge: Harvard University Press, 1984.

__________. Self-Command in Practice, in Policy, and in a Theory of Rational Choice, *American Economic Review: Papers and Proceedings,* May 1984, 1-11.

__________. Coping Rationally with Lapses from Rationality. *Eastern Economic Journal,* 1996, 251-69.

**Sen, A. K.** Choice, Orderings, and Morality, in *Practical Reason,* edited by S. Korner. Oxford: Blackwell, 1974, 54-67.

__________. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs,* 1977, 317-44.

**Thaler, R. and Shefrin, N. M.** An Economic Theory of Self-Control. *Journal of Political Economy,* April 1981, 345-74.

**Van der Veen, R. J.** Meta-Rankings and Collective Optimality. *Social Science Information,* 1981, 345-74.

**Walsh, V.** *Rationality, Allocation, and Reproduction.* Oxford: Clarendon Press, 1996.

**Weisbrod, B. A.** Comparing Utility Functions in Efficiency Terms or, What Kind of Utility Do We Want? *American Economic Review,* December 1977, 991-95.