Preferences, Metapreferences, and Morality

BENJAMIN MARROW
Yale University

THAT humans have certain desires is a fundamental truth of our nature and a central premise of economic theory. One may disagree about what those desires comprise or whether such desires are morally good or even beneficial for our interests, but one cannot doubt the existence of preferences across choices and alternatives. Economists, rational choice theorists, and many political scientists operate under the assumption that individual agents in the market and in the community have (somewhat) stable preferences and that individuals act according to these preferences.

That we have desires *about* desires is far less clear, but as important a concept. As Plato notably discussed through his concept of *akrasia*[1], the process of acting against one's better judgment, there seems to exist a disconnect between one's desires and one's preferences about those desires. We all have desires that we desire not to have; at the same time, we often wish that certain beneficial activities constituted our desires.

Despite the apparent implications for the social sciences, the topic of metapreferences heretofore has been discussed primarily within the realm of philosophy. Philosophers of mind have discussed the concept of metapreferences as it relates to the will and human autonomy.[2] According to certain philosophers, the ability to evaluate our own preferences and to act against our first order preferences is what differentiates humans from non-humans and renders us free with respect to our will. In social sciences, meanwhile, much of the discussion of metapreferences has focused on specific explanatory applications—for example, on reasons why people commit suicide or on questions of social choice stability—or on reasons for why traditional economic preference theory may be insufficient in modeling human behavior.[3] Scholarship regarding the application of metapreferences in the social sciences has been sparse, and those who have discussed it have not fully considered the objects or content of our metapreferences.

In this paper, I will examine the formation and application of metapreferences to

1 Plato. *Protagoras*. In *Plato Complete Works*. Ed Cooper, John M. (Indianapolis: Hackett). 358d.
2 See for example Harry Frankfurt. "Freedom of the Will and the Concept of a Person". *The Journal of Philosophy,* Vol 68, No. 1 (1971); Richard C. Jeffrey. "Preferences Among Preferences". *Journal of Philosophy* 71 (13):377-391 (1974); or Gerald Dworkin. *The Theory and Concept of Autonomy*. (Cambridge: Cambridge University Press, 1988). As Dworkin, writes, "Autonomy is conceived of as a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher order preferences and values".
3 See Albert O. Hirschman. "Against Parsimony: Three Ways of Complicating Economic Discourse" in *Rival Views of Market Society and Other Recent Essays*. (Cambridge: Harvard University Press, 1986).

argue that metapreferences fall into their own distinct, conceptual category. Secondly, I argue that the content of metapreferences draws on a variety of society-based and value-based heuristics. In particular, by invoking heuristics of popular approval and individual moral intuitions, metapreferences lead us to a set of preferences less selfish than what has been thought but ones that nonetheless cater to our wellbeing.

What is a Metapreference?

Given the complexity of the concept of metapreferences, there is no singular accepted definition that satisfactorily encompasses the multiplicity of applications across different fields. Broadly speaking, however, a metapreference takes the general form of an aversion to or an approval of one's desire for a specific good or activity. In more formal logical notation, an aversion to one's desire might be expressed as *A pref. X over Y* but *A pref. [A pref. Y over X]* to *[A pref. X over Y]]*. In other words, A wants X over Y, but A wants to want Y over X. Or as Frankfurt puts it, "besides wanting and choosing and being moved *to do* this or that, men may also want to have (or not to have) certain desires and motives."[4]

The possibilities for metapreferences are numerous, but perhaps the most obvious and illustrative manifestation of a metapreference is one in which an individual has a preference for a "criticizable" activity.[5] A characteristic example would be that of the smoker, who desires a cigarette, but who wishes she did not have that desire. This individual is said to have a first-order desire for a cigarette, but a second-order desire (metapreference) that expresses distaste for her first-order desire to smoke. As Hirschman notes, metapreferences arise from the "ability [of humans] to step back from their 'revealed' wants, preferences" and evaluate them accordingly.[6] It is only through "stepping back" and evaluating the desire that the smoker can realize a metapreference, for she surely does not have a first-order desire not to smoke.

Before continuing, it is important to differentiate between several different conceptions of metapreferences. The primary conception of metapreferences I address in this paper supposes that "wanting" a certain desire connotes "wanting [the desire] to guide what I ultimately choose." In other words, my metapreference for desire Y over desire X entails that I want the end of desire Y. In many ways, this formula highlights the semantic difference between a "want" and a "preference," though as Frankfurt notes "it could not be true both that A wants the desire to X to move him into action and that he does not want to X."[7]

One could, however, imagine examples in which an individual could want to have a certain desire purely to experience the desire itself, without regard for its "end." This might be called the qualia view of metapreferences, and its occurrence and practical significance is harder to imagine. One can think, for example, of a therapist who wishes to want drugs in order to understand what desires his patients experience; Odysseus

4 Frankfurt, 1971, 7.

5 Later in the paper I will discuss exactly what is meant by 'criticizable'.

6 Hirschman, 1986, 144

7 Frankfurt, 1971, 10. In other words, a preference could be thought as a more holistic evaluation of alternatives that includes both the strength of the desire and a rational evaluation of its end, whereas a want might simply include a desire.

wishing to experience the temptation of the Sirens[8]; or an ascetic monk who does not want to want money (but as Douglas Whitman notes, would be fine with it appearing on his doorstep)[9]—but it is rare that we look at desires qua desires.

A more common type of metapreference, which might be termed the virtue-ethical metapreference, would be the act of wanting to be the sort of person that wants Y—for example, an individual might give to charity because he wants to be the sort of individual who engages in philanthropic activities. While this could be a type of primitive intuition, presumably this type of intuition could also appeal to either the desire itself or to the end of that desire. Indeed most of the time we wish to have them in order to become the kind of person who acts on those desires.

Bernard Grofman and Carole Uhlaner introduce a fourth conception of metapreferences. For them, metapreferences are "preferences over characteristics of choice processes," or "preferences for the features of the procedures which result in outcomes, and not simply preferences for outcomes *per se*."[10] Grofman holds that metapreferences concern the mechanisms involved in making choices, and suggests that the framing of the choice procedure can affect our first-order preferences themselves. In other words, given that preferences are, by nature, choices among alternatives, the rules of the choice process can be a source of a metapreference. To be sure, this definition is different from the iterative "second-order" and deliberatively judgmental nature of metapreferences discussed by such scholars as Frankfurt, Dworkin, and Hirschman, but it parallels their view by framing how we act.

Despite these differences in definitions, it is important to note the similarities. All four conceptions suggest that the current rational choice model that only evaluates the revealed first-order desires of individuals (i.e. one's preference is simply what one chooses to do) is fundamentally flawed. First-order desires do not always direct our actions, and each definition implies that a deeper, underlying set of preferences ought to be considered in concert with revealed desires. Secondly, as I will explore in the final section of this paper, there is significant overlap in the content that characterizes the different sets of metapreferences, including factors that are not immediately self-interested. For the purposes of this paper, I will take a metapreference to be an end-oriented heuristic used for evaluating desires that is neither self-interested nor is driven by concerns about utility.

Debates over Metapreferences

Much of the resistance to the inclusion of metapreferences in contemporary economic literature stems either from a refusal to accept the existence of metapreferences as a distinct category or a failure to acknowledge their practical importance. Many economists and rational choice theorists hold that the satisfaction of any preference, by dint of it being a preference, brings utility. On this view, metapreferences are just a variant form of first-order preference. The economist Gary Becker, for example,

8 Jon Elster, "Ulysses and The Sirens: A Theory of Imperfect Rationality". *Social Science Information* (5). 469.

9 Douglas G. Whitman. "Meta-Preferences and Multiple Selves". (California State University, 2003).

10 Bernard Grofman and Carole Uhlaner. "Metapreferences and the Reasons for Stability in Social Choice: Thoughts on Broadening and Clarifying the Debate". *Theory and Decision* 19 (1985) 31.

acknowledges that "economists have had little to contribute … to the understanding of how preferences are formed," and so under the "economic approach," "preferences are assumed not to change substantially over time, nor to be very different between wealthy and poor persons, or even between persons in different societies and cultures."[11] He further constrains preferences, arguing that not only are they assumed to be stable but that they are also "defined over fundamental aspects of life, such as health, prestige, sensual pleasure, benevolence, or envy."[12] There is no room for "ad hoc shifts in values," and whatever "non-rational" behavior does occur can be explained according to incomplete information or "the existence of costs, monetary or psychic."[13]

While Becker does not address the concept of metapreferences directly, it is easy to see why the concept of metapreferences I discussed previously could not cohere with this view. As Hirschman notes, metapreferences often *only* come to light through conscious changes in action (or in economic terms, changes in revealed preference).[14] This is due to the fact that if one's metapreferences always corroborate one's first-order preferences, then metapreferences have no practical significance. Rather, "certainty about the existence of metapreferences can only be gained…[in] changes in actual choice behavior;" namely when one's metapreference is directly at odds with her preference and causes her to act differently.[15] In the example of the smoker, we can gain insight into metapreferences precisely through the observation that on one day she refuses to smoke (i.e. when her revealed preference has changed), despite the fact that there is no observed change other than her taking the time to self-evaluate. To be clear, this is not to say that all changes in preference are motivated by metapreferences, nor that all metapreferences necessarily cause changes in behavior. It is possible that a metapreference could disagree with a first-order preference (that is to say, the conflict of interest exists), but the metapreference is weaker than that preference. In this case, the existence of conflicting preferences would detract from overall utility, but the metapreference would not have sufficient power to effect a change in action. Rather, it seems that there is a relevant sub-category of metapreferences that can effect changes in behavior.

Douglas Whitman, an economist who acknowledges several of the shortcomings associated with the overly rational approach pioneered by Becker (Whitman presents his own theory to explain anomalies of choice) is similarly skeptical of the concept of metapreferences. According to Whitman, and consistent with Becker's approach, a second-order desire often stems from a "frustrated" first-order desire.[16] To return to the case of the smoker, Becker and Whitman would argue that the reason he stops smoking is that he has both a first-order desire for health and a first-order desire for cigarettes. When his desire for cigarettes "frustrates" another first-order desire, he inevitably gains a metapreference. In this sense, Whitman understands what we call metapreferences to be simply "intellectualized" desires.[17]

There is certainly some value to this view of metapreferences that ought be

11 Gary S. Becker. *The Economic Approach to Human Behavior*. (Chicago: University of Chicago, 1976), 5.
12 Ibid.
13 Ibid., 6.
14 Hirschman, 1986, 144
15 Hirschmann, 144.
16 Whitman, 2003, 5.
17 Ibid., 12.

considered. Importantly, Whitman (and Becker) highlight a common point of confusion in current discussions about metapreferences. It is not simply that in rejecting a cigarette, one demonstrates a metapreference—after all, as Becker notes, health is a commodity (it brings us utility) and many of us who reject a cigarette may be contradicting an "instinct,"[18] but we are ultimately doing so to serve our health. But Whitman and Becker's discussions overlook two important differences. The first is the difference between tastes and values, and the second is post-hoc theory development.

Frankfurt's description of metapreferences in *Freedom of the Will and Concept of the Person* does not explicitly mention the distinction between tastes and values, but Frankfurt implies it through his distinction between "persons" and "wantons." Frankfurt's conception of personhood relies on one's capacity to see the "desirability of his desires," which even a rational wanton—an individual who always acts to maximize her utility—could not do.[19] What exactly does Frankfurt mean by this? Consider Whitman and Becker's idea that metapreferences are simply another form of preferences and tastes. If this were the case then humans would *only* be concerned with maximizing utility, rather than having an underlying preference for the desire. This would mean that the smoker would simply be performing a cost-benefit calculation without caring which of his conflicting desires would take precedence in the end. Yet this conclusion appears tenuous at best. It seems that the smoker does not refuse the cigarette merely to maximize her utility according to some intertemporal utility judgment, but rather has underlying values about the activity itself and her relationship to that activity which transcend the activity's measure of utility. In other words, Whitman and Becker seem to suggest that, so long as two activities bring an individual the same utility, that individual would have no reason to prefer one activity to the other.

The argument that Whitman and Becker might put forth in response to this criticism—that we wouldn't choose something if it didn't maximize our utility, or that there must be some implicit cost that has been overlooked—may certainly be plausible but rests on tautology or post-hoc theorizing. As Green and Shapiro note, rational choice theorists will often approach questions by "engag[ing] in a thought experiment designed to generate an explanation of a given phenomenon that is consistent with rational choice assumptions, somehow specified."[20] This issue is further complicated by the fact that "the predictions of one rational choice model will invariably overlap with those derived from another kind of theory."[21] Simply because refusing a cigarette seems to benefit our utility, it does not follow that we make that choice so as to maximize our utility. Indeed, as I will argue, our metapreferences are invariably more moral in character, which can benefit our self-interest (e.g. self-interest well understood) but not *because* we are self-interested.

Even if one were to entertain Whitman's argument that metapreferences and

18 While it is very often the case that the metapreference-preference distinction will parallel a reflective-instinctual difference, this is not always the case. As is later discussed, certain metapreferences—such as those concerning popular support or procedural fairness—can be instinctual in nature. Similarly, one can also take the time to reflect on what one really desires.

19 Frankfurt, 1971, 11.

20 Donald P. Green and Ian Shapiro. *Pathologies of Rational Choice Theory*. (New Haven: Yale University Press, 1994). 34.

21 Ibid., 37.

preferences impact the same utility (one that takes into account all manner of psychic and monetary benefits), it seems there is still something to be said for understanding second-order desires as a distinct category. In labeling metapreferences "intellectualized desires," Whitman implicitly acknowledges that metapreferences are of a different nature. For example, "meta-preferences can operate under specific circumstances, such as when an individual binds himself in advance;"[22] second-order desires more often than not "demand psychic attention" (as opposed to material satisfaction);[23] and our ideas of metapreferences often concern our "welfare utility function" (what is good for us) while first order preferences concern our "behavioral utility function" (what we want).[24] Thus even if we *can* compare the utility of satisfying a preference and satisfying a metapreference, if the formation, application, and content of the two are different (which Whitman suggests), perhaps each ought to be considered in its own scope.

Metapreferences in Practice

To the extent that we have examined metapreferences in theory, it will be worth looking at what pragmatic implications the theory of metapreferences provides, particularly as they concern the economist or political scientist. After all, one can endorse Frankfurt and Dworkin's arguments for the existence of second-order desires and their role in establishing human autonomy without believing that second-order desires hold any serious consequences for policy or action. Yet as Hirschman argues, metapreferences are valuable precisely *because* they reflect our values (that is, he argues, values are a source of our metapreferences) and allow us to act against our preferences or "interests." We have already considered what this looks like in one situation—namely, when someone rejects a cigarette because he does not want to want to smoke—but no scholar has properly considered how metapreferences are manifested on a larger scale.[25] I do not claim to be able to fill this gap with empirically validated examples, but I would like to suggest some situations (albeit post-hoc ones) in which metapreferences might be at play. These situations provide room for exploration in further studies.

At one level, metapreferences can explain why we engage in restrictive behavior. People, by and large, do not adhere strictly to a preferentist model of behavior, and will often take steps in advance to remove criticizable goods from their path. At the same time, individuals can make themselves do what they do not "want" to do. Consider the voter's paradox, which, as Green and Shapiro demonstrate, is a problem with rational choice theory. There appears to be no valid self-interested reason as to why individuals want to go to the polls either from a welfare standpoint (it is not a sacrifice from which they gain future benefits) nor from a behavior standpoint (the present-time "consumptive" benefits from attending the polls are virtually non-existent).[26]

22 Whitman, 2003, 7.
23 Ibid., 15
24 Ibid., 2
25 Hirschman's discussion of metapreferences indicates that there are pragmatic consequences for understanding human behavior, without detailing what those consequences would look like.
26 Green and Shapiro, 53.

Given that there is no obvious benefit whatsoever from attending the polls, the voter's paradox addresses why so many still vote. A metapreferential model might be seen to provide some clarity in this regard: we don't have a preference for voting per se, but we might want to be the type of person who is civically minded and so wants to vote, and so we make ourselves go to the polls. These effects gain in clarity when we look at some empirically demonstrated examples, particularly ones where there is a conflict between moral forces (our values) and our tastes (our first-order desires). One famous case is that of the day-care center in Haifa Israel that saw a dramatic increase in parents picking up their children late, *after* the center started charging parents for picking up children later.[27] This runs directly counter to Becker's suppositions about demand. Certainly there could be a Beckerian explanation to this change in behavior that examines the change in terms of latent psychic costs that, once the policy was implemented, undermined an existing disincentive to keep children late, but the change in behavior could also be potentially explained by the idea that individuals did not want to be the type of person who have a revealed preference for leaving children late (their metapreference prevented them from leaving children late more often). It is only once the charge was instituted that the parents' metapreference was satisfied at the expense of their first-order preference for money. There are many comparable examples in economic literature of cases where economic factors "crowd out" moral considerations, suggesting that traditional economic models of human behavior are incomplete.

Another key practical aspect of metapreferences concerns the formation of metapreferences. Hirschman touches on the importance of "stepping back from" desires, which suggests a more reflective and less instinctual process. Attempts to influence our preferences often deal with one of these approaches at the expense of the other. For example "advertising and other acts of marketing influence the preferences that agents experience but do not influence the metapreference ranking."[28] This formation is consistent with the taste-value distinction made earlier: advertising works to formulate our tastes (for example, flashing an unhealthy food before the screen) in an attempt to ensure our first-order preference takes priority. A reflective or meditated process, wherein we take time to evaluate our desires and wants, meanwhile, would seem to favor the instantiation of our metapreferences over our first-order preferences, due to the fact that many individuals do not like the fact that they want, say, an unhealthy food. In these situations, where we are unable to actualize our metapreference as our will, Frankfurt might argue we lose our conception of personhood and simply become rational wantons responding to utility without considering the desirability of our preferences. Adam Smith also highlights this aspect of metapreference formation when he suggests that "the eagerness of passion will seldom allow us to consider what we are doing with the candour of an indifferent person" and it is only when are not seized by passions that "we can enter more coolly

27 Samuel Bowles, "Policies Designed for Self-Interested Citizens May Undermine the 'Moral Sentiments': Evidence from Economic Experiments," *Science*. Vol. 230 No. 5883 (June 20, 2008), 1605-1609.

28 David George, "Does the Market Create Preferred Preferences?" *Review of Social Economy,* Vol. 51, No. 3 (Fall, 1993), 333.

into the sentiments of the indifferent spectator."[29]

While these areas ought to be explored further, some discussion of metapreferences has already pointed to explicit situations in which our metapreferences do determine certain non-rational behaviors. Grofman's discussion of metapreferences, for example, is part of an attempt to explain social choice stability—that is, why there is often "too much stability in social choice processes," despite the fact that some of these outcomes might be less than favorable.[30] David George and David Lester, meanwhile, have shown empirically how metapreferences affect suicidal tendencies by demonstrating that "people whose metapreference is for life over death may be less at risk for suicide than those who metapreference is for death over life" and that it is therefore impractical to simply consider first-order preferences.[31] George also suggests that understanding metapreferences is necessary for a normative analysis of economic institutions, and not merely to "render a richer model of human choice."[32] According to this line of reasoning, even if the economist does reject the idea that we can use metapreferences to direct our actions, economic institutions based purely off tastes that ignore the harms to welfare and metapreference satisfaction hardly maximize utility at all. In such a treatment of humans—one based on tastes and not values—"the market displays a pervasive inefficiency in its preference producing capacity."[33]

Metapreferences and Virtue

Discourse on metapreferences has also been conspicuously devoid of a thorough discussion of what exactly metapreferences consist of and what they satisfy, if not utility. Frankfurt, for example, reflects on the "suitability" of desires without explaining what is meant by suitability; he posits that there is no clear answer, stating that "there is no essential restriction on the kind of basis, if any, upon which [metapreferences] are formed."[34] Grofman gives the most expansive outline of what constitutes our metapreferences, but similarly remains agnostic as to what extent these metapreferences are expansions of rational choice. In his view, metapreferences include such things as "procedural fairness," "consensus," "universalism and civility norms," and "preference for decision maker's image."[35] Examples explored in other scholarly papers seem to hint at, without explicitly stating, similar types of standards for metapreferences, and in particular, ones that provide for our welfare. In this section, I suggest that the pertinent context for second-order preferences includes such considerations as morality (a first-order desire is criticizable if it does not cohere with our ethical intuitions) and relatedly, societal values (a first-order desire is criticizable if others would not approve of it).

One understanding of how these considerations affect the formation of metapreferences can be seen deductively. Consider an individual who takes the time to step back and evaluate her preferences. By definition, the individual cannot appeal

29 Adam Smith. *The Theory of Moral Sentiments.* (Oxford: Clarendon Press, 1976). 157.
30 Grofman and Uhlaner, 31.
31 David Lester and David George, "Metapreferences, Preferences and Suicide: Second Column in a Series," *Crisis*, Vol. 21: 2. (2000). 57-58.
32 George, 1993, 332.
33 Ibid., 344.
34 Frankfurt, 1971, 13.
35 Grofman and Uhlaner, 1985, 40-44.

to a first-order desire: if the standard by which the individual evaluated her desire was the extent to which it cohered with her first-order desires, then her metapreference could not contradict her first-order preference, and would hardly be a metapreference at all.[36] Another related possibility is that the metapreference could be associated with a facet of human nature such as "risk aversion" that does not exactly constitute a "desire" but neither is a "standard" reached upon reflection. For example, Grofman places an emphasis on "uncertainty avoidance" as a common metapreference, wherein "decision makers may place a high value upon maintenance of existing decision-making institutions and procedures, for reasons which may include custom and uncertainty avoidance."[37] Frankfurt also admits the possibility of instinctual but not desire-fulfilling metapreferences, such as when "a person may be capricious and irresponsible in forming his second-order volitions and give no serious consideration to what is at stake."[38] Past these minor considerations, however, a person who crafts a deliberate second-order volition must turn to other standards.

Adam Smith identifies several of these standards. In his *Theory of Moral Sentiments*, Smith describes some mechanisms that could serve as standards for metapreferences, such as his suggestion that we should place our desires and our actions in the context of the type of person others would approve. According to Smith, the "moral sentiment" in humans is primarily a means by which we evaluate others, and the extent to which others would approve of an action is a litmus test for the action's suitability. He explains how things are "regarded as decent, or indecent, just in proportion as mankind are more or less disposed to sympathize with them."[39] Smith's idea parallels that of Grofman's description of the metapreference of the importance of "consensus," or as he puts it, the question of whether "the outcome[s] have (or appear to have) substantial popular support?"[40] But while Grofman admits this factor "could be rational if the cohesion facilitates obtaining other valued goods," Smith understands that the metapreference transcends standard measures of utility, and rather serves as a heuristic for what is good: it is not that we actually seek approval from others, but rather that we desire "not only to be loved, but to be lovely; or to be that thing which is the natural and proper object of love."[41] An individual has an underlying preference that exists "independent of any advantage which he can derive from it."[42] Smith continues, "Nature, accordingly, has endowed [man], not only with a desire of being approved of, but with a desire of being what ought to be approved of being what he himself approves of in other men."[43] One implication of defaulting to others is that our metapreferences are in large part shaped by the values of the society in which we live. We may have a metapreference against certain non-conforming sexual preferences not so much because of utility or morality, but simply because certain activities are

36 A virtue ethicist like Kelly Rogers might argue serving our welfare is the "right" thing to do and so like Robinson Crusoe on an Island, we "ought" to provide for our welfare but not simply to increase our utility and not because we are selfish. See Kelly Rogers "Beyond Self and Other." *Social Philosophy and Philosophy* 14 No. 1 (1997).
37 Grofman, 1985, 41
38 Frankfurt, 1971, 13.
39 Smith, II.i.2.2.
40 Grofman, 41.
41 Smith, III.ii.i.
42 Ibid., III.i.6.
43 Smith, III.ii.2.

frowned upon in our society. While this suggests that metapreferences are unstable, there is perhaps something to be said for the aggregate judgment of mankind. At least according to Smith, human approval corresponds closely with the justice of an action, and so social approbation and moral intuitions are hardly ever distinct.

Smith also introduces another standard by which we make evaluations—that of the third-party observer. In many ways, when we take the time to step back and reflect on our desires, we take on the role of a third party observer. For Smith, this mechanism of evaluation holds significant normative implications, for it is only when we "call forth…the impartial spectator" and look from the "place and eyes of a third person" that we can ignore our desires and "judge with impartiality between us [and others]."[44] In this type of evaluation therefore, man can properly "humble the arrogance of his self love, and bring it down to something which other men can go along with."[45] As mentioned, there is a heavy overlap for Smith among what type of person we desire to be, what other people approve of, and what is morally good.[46] Even apart from other people's approval, and from the impartial spectator, both of which lead us to moral ends in Smith's view, there is the aforementioned mechanism of coming to our decisions through a reflective and reasoned process. Given that it is only through "reason, principle, conscience" that we are "capable of counteracting the strongest impulses of self-love," it seems that metapreferences, which are more reflective in nature, will lead us to a slightly more moral sense of self-interest.[47]

Smith's discussion of self-evaluation and reflection mirrors the empirical facts. Almost all of Grofman's proposed metapreferences adhere to maxims of morality or of widespread approval. Procedural fairness and universalism norms, for example, point to the role of impartiality in forming our metapreferences, while questions of civility norms and consensus emphasize the role of societal approval. Similarly, in looking at the examples of metapreference-preference conflicts, we rarely see an example of an "immoral" or widely discouraged metapreference. We desire not to want the cigarette or unhealthy food; we desire to want to study hard and not be greedy. Our metapreferences, in this sense, cause us to be good and esteemed people and guide our self-interest as such.

Conclusion

The argument put forth in this paper is bold, if only because of the epistemological difficulty (and perhaps even impossibility) in proving the existence of metapreferences. If anything, the primary purpose of this paper is to suggest the need for further studies of a nascent but significant sphere of preference theory. What I have sought to do in this paper is to sketch a preliminary synthesis of existing ideas about preferences in order to see how a "philosophical" concept might apply to our understanding of rational choice and self-interest. It seems evident that current models of preference theory are insufficient. Regardless of how one defines preferences, because we have the

44 Ibid., I.i.5.4.
45 Ibid., II.ii.2.1.
46 Smith argues that the "perfection of human nature" lies in loving ourselves only as much as we love each other, and suggests that this is the most agreeable sentiment to mankind (I.i.5.5).
47 Ibid., III.iii.5

power to evaluate ourselves and direct our actions accordingly, simply looking at what we "want" on the first-order level can obfuscate a dynamic view of human behavior. In particular, policies can be more effective if they take into account both the manner in which we form our metapreferences and determine their content. Individuals have desires, but they also evaluate those desires according to preferences seen to be good by others. In this sense, metapreferences can be understood as underlying matrices for human behavior: they ground our first-order preferences in moral structures which sustain our vision of a society that is simultaneously just and based on self-interest. ■