PREFERENCES AND THE EXPLANATION OF SOCIAL BEHAVIOR

Jeremy Freese

Department of Sociology

Northwestern University

jfreese@northwestern.edu

Word count: 9,122 including references and footnotes

DRAFT (August 25, 2008). To appear in the *Oxford Handbook of Analytic Sociology*, edited by Peter Hedström and Peter Bearman.

Analytic social science endeavors to explicate iterative connections between the properties of a social system and the action of individuals. In a celebrated example from Schelling (1978), patterns of racial segregation are connected to an individuals' either staying where they live or moving (see also Bruch and Mare, this volume). Such projects require a logic of how actors respond to circumstances. In Schelling's model, actors moved if the proportion of neighbors "unlike" themselves exceeded a threshold. This can be expressed as a simple rule ("if more than two-thirds of neighbors are unlike oneself, move; otherwise, stay"), and analytic work might simply model actor behavior as following rules. Yet, our appreciation of the exercise as telling us something instructive about human affairs is greatly enhanced if the behaviors implied by the rules can be interpreted in a way that makes sense to us as human action (Hedström 2005). Schelling's model becomes intelligible as a model of meaningful, understandable action when we suppose that individuals are *deciding* whether to stay or move, and these decisions are influenced by their *wanting* at least a specified proportion of neighbors to be like themselves.

Understanding behavior as *intentional* pervades everyday life. We experience ourselves as wanting things and acting on those wants. We apprehend others as having wants of their own, sometimes overlapping with ours and sometimes not, and our beliefs about what others want influence our interactions with them. Our sense of people as pursuing purposes feels "natural" and works reasonably well in everyday life, and social scientists do not leave this sense behind when we come to the office. Social science does not just use intentional interpretation in giving meaning to agents in simulations like Schelling's, but also in explaining and predicting action. For example, toward explaining the observed network of romantic relationships in a high school, Bearman, Moody, and

Stovel (2004) invoke a model in which adolescents want partners with characteristics like themselves and want to avoid negative implications that might follow from dating an ex-partner's current partner's ex-partner.

As commonplace as it is to talk about wants, reflection quickly indicates that their incorporation into social scientific explanation is not straightforward. The wants of others are not directly observable; instead, we often seem to infer them from the very phenomena (their behavior) that we are otherwise trying to explain. Moreover, we know that we often feel uncertain or ambivalent about what we want, that we can feel we want something but not act like it, and that we are not always honest to others about what we believe we want. How, then, should social scientists think about the place of wants in explaining action? How do wants fit into efforts to explain large-scale social phenomena partly by reference to the actions of individuals?

Social scientists use a constellation of terms to characterize wants and closely related counterparts, including desires, tastes, preferences, goals, values, aspirations, purposes, objectives, appetites, needs, motives, ends, drives, attractions, orientations, and wishes. Of these, analytic social science has made the most use of "preferences," and so "preferences" serves as the focal term for this essay. The preference for "preference" stems partly from "preference" implying alternatives. We desire things in their own right, but we prefer things to other things. We can desire both X and Y and still prefer X to Y. We also can desire neither X nor Y and still prefer X to Y. Preferences thus become an especially sensible term of art if one regards the world as full of trade-offs and actors as wanting many things (and a pony, too!) but only being able to obtain some of them.

The concept of preferences can be tightened with assumptions to yield rational actor models in which actions are posited to be optimal. Rational actor models dominate economic thinking, but they have been used less in sociology. "Preference" here is not meant to presume invocation of rational actor theory. Instead, the relationship between "preference" generally and its specific usage in rational choice theories will be recurrently considered in what follows. More broadly, beyond its appropriateness to comparison of undesired alternatives, I do not make any working distinction in this essay between "preferring" X to Y and "desiring more" or "wanting more." Certainly, "desire" is often used to convey more the experience of wanting, and "values" are regularly intended to orient attention to basic moral and related wants, but these distinctions are not my focus here.

This essay considers four basic issues regarding "preferences" as an explanatory concept in analytic sociology. First, I take up how the ontology of preference should be understood, that is, the question of what preferences are. I argue against both that preferences are "mental events" or are "behavioral tendencies." Second, given that real-world alternatives may be characterized by numerous different attributes, I consider how to approach the question which attributes may be most salient to understanding action. Third, I discuss applying "preferences" to organizations, and, while I contend there is no principled reason why preferences are more naturally a property of persons than of organizations, I maintain that the orientation of sociology to the preferences of each is rightly different. Finally, I discuss preference change over time, with especial focus on the preferences of actors being a target of purposive action by others and by actors themselves.

WHAT ARE PREFERENCES?

For purposes here, *explaining* action is understood as constructing narratives that make the occurrence of action *y* when and how and by whom it occurs intelligible to an audience by reference to causes. *Causes* of *y* are events and states of the world for which, if different at the pertinent point in time, we assert that (the probability distribution for) *y* would be different.[1] *Immediate causes* of *y* are causes whose place in the explanation is narratively contiguous to *y*. Immediate causes are of course themselves caused, and causes of causes are caused, and so on. Explaining particular actions is thereby an indefinite project, in which the narrative network may extend *backward* in time, *inward* to the psychology/physiology/neurology of actors, and *outward* to the social and physical environments in which actions are determined (Freese forthcoming). Explanations are never "full" or "complete" in any essential sense but rather at best adequate to the practical purpose of providing understanding for some audience.

From this perspective, preferences may be understood in the first instance as being among the prototypic immediate causes of actions understood as intentional. Someone offered vanilla or chocolate ice cream chooses vanilla. "Why?" "She prefers vanilla." This is not an empty statement given possible alternatives: "She actually prefers chocolate, but that chocolate scoop is already half melted." As explanations go, the statement suffices in

[1] Hedström (2005: 39) defines desires as mental *events*, perhaps because he wants to consider desires as causes and recognizes that many philosophers quarrel with causes that are not events (p. 14). Yet, social science invocations of "desire" and "preference" in explanations—including the examples in Hedström's book—refer overwhelmingly to continuing conditions or states of actors. Either some term other than "cause" needs to be available for discussing the counterfactual implications of different states, or social science needs to resist the idea that only events are causes.

the same sense in which Boudon (1998) says that rational action is its own explanation. Given this preference and a set of background conditions, the action *follows*, and so the explanation provides a causal narrative that is logically continuous in the sense that there is no narrative gap intervening between *explanans* and *explanandum*. Offering "She prefers vanilla" as sufficient suggests "Under the circumstances, had she preferred chocolate, she would have chosen chocolate." The psychology of the actor is asserted to be decisive, and "preference" is the way in which the decisive difference is characterized. While sufficient in this sense, the explanation might be regarded as (obviously) practically deficient, in that we may feel inadequately informed without knowing more about why the person prefers vanilla to chocolate or why the person came to be offered chocolate and vanilla instead of some different set of flavors or no choice at all.

What does it mean to say someone prefers vanilla to chocolate? Clearly, preferences are deployed in explanations as attributes of actors, as something actors *have*. One possibility is that preferences may be identified with some physically real characteristic of brains. Maybe to say someone prefers vanilla to chocolate is to say that a prototypic vanilla scoop stimulates a stronger dopamine reward response in the brain (or whatever) than a prototypic chocolate scoop. While one hope of "neuroeconomics" is lend insights into the internal processes involved in choice comparisons (Camerer, Loewenstein, and Prelec 2005), a large amount of philosophical and cognitive science argument provides abundant reason to doubt that preferences (or desires abstractly) can be reduced to any kind of physical definition (see Dennett 1987; Ross 2005). I will not rehearse the arguments here, but simple consider that it is unlikely that all people who prefer chocolate to vanilla could be said to do so for the same reason, or that all manner of preferences (amongst foods, jobs, leisure activities, romantic partners) could be reduced to the same internal differentiating process. Indeed, preferring vanilla to chocolate does not even imply that a person experiences vanilla as better tasting, as one might think chocolate tastes better yet prefer vanilla for other reasons (e.g., a belief that vanilla is healthier, an aversion to brown foods).

A different possibility is that preferences are characterizations of a behavioral pattern or tendency, so that "I prefer vanilla" is just another way of saying "I tend to choose vanilla." In addition to possibly conflate attributes of actors with those of situations, identifying preferences as behavioral patterns undermines their status as cause and thus their relevance to explanation. Saying "I tend to choose vanilla" articulates a pertinent statistical regularity, but a statistical regularity in itself is not a cause (but possibly reflects a common underlying cause that produces the regularity). Revealed preference theory in rational actor theory takes actions as manifestations of a transitive ordering of preferences over all outcomes (Samuelson 1947). In some interpretations (e.g., Gintis 2007), this implies preferences are part of a re-description of action and so no more explanatory than "I chose vanilla because I chose vanilla." Clearly, either revealed preference theory or that interpretation of it must be rejected if preferences are to remain part of one's explanatory vocabulary. Nonetheless, reference to recurrent choice and behavioral patterns seems *close* to what preferences might be, especially since considering past actions would seem to provide the plainest grounds on which preferences may be inferred.

The solution might seem to regard preference as a disposition to choose one alternative over another, but this raises the question of just what we mean by "disposition." Clearer would be to consider preference as making reference to the alternative that would be selected in *a counterfactual situation of abstract, hypothetical choice*. To say an actor prefers vanilla to chocolate is to say that either the actor will choose vanilla over chocolate or there will be some reason(s)—drawing upon other preferences, beliefs, or circumstances—the actor does not. A statement of an abstract preference of X over Y bears affinities to an idea of a "default setting" in computer science or an "unmarked form" in linguistics. Preferences thereby specify a default expectation about choice that may be part of a sufficient immediate explanation when it is consistent with observed choice and prompt further explanatory elaboration when it is not.

In this view, preferences are not to be reduced to any specific property inside an actor's head. The preferences of individuals may appear to be "mental states," but they are not mental states in the sense that we should imagine neuroscience or any other enterprise to eventually be able to describe specific circumstances of the brain in a way that maps unproblematically on to what social science means when it says an individual prefers X to Y. Instead, preferences are *holistic simplifications* of the internal workings of actors that are used when making sense of behavior as intentional. Put another way, when we interpret action as intentional, we take the actor to be a system, and preferences are attributed as a property of the operations of that system treated as a whole (Dennett 1987). Philosophers often refer to preferences as part of our "folk psychology," but better for understanding their explanatory role in social science is to understand them as *black-boxing psychology*, as simplifying the messy and mysterious internal workings of actors. Preferences accomplish this simplification by representing inner complexity in the most immediately action-implicative terms, by distilling complexity through the bottleneck provided by hypothetical decision. Whatever goes on in the brains of actors, the question of whether they prefer chocolate or vanilla comes down to the resulting implication for our expectation about a hypothetical choice. (Although outside the purview of this essay, one can take a similar stance regarding beliefs, considering them as propositions about what hypothetical choices imply for what an intentional actor apparently regards as true. [Ryndgren, this volume, adopts a more introspective and cognitivist perspective on belief that fully decouples it from choice and action])

Preferences and beliefs thereby serve as a model to whatever complicated machinery produces the action of the systems to which they are applied. Even so, if the model is to work in any kind of consistent way, attributes of preferences characterize real, internal operations of the actor, even though only at the emergent level of choices. To be sure, understanding exactly what is going on internally to yield a particular preference takes us to a more explicitly cognitive or neurological vocabulary and may be wildly different for different actors. But if we reduce preference only to its usefulness for predictive claims, we lose it as a matter of causal claim and thereby as a term for explanation. Preferences can be used in explanation so long as we recognize that we are understanding ourselves as using a simplification that characterizes holistically a more complicated (and multiply realizable) underlying reality. When we say a preference caused an action, we are asserting that if the internal operations were such that an alternative characterization of the preference would instead apply, then the action would be different.

Preferences, along with beliefs and the rest of the intentional idiom, stands as both partition and interface between analytic social science and psychology. Preferences are useful for efficient, parsimonious characterization of the behavior of what Coleman (1990) calls “natural persons.” Schelling’s model implies psychological processes that yield a preference to have at least some neighbors like oneself, but we do not have to know what these processes are to render the action in his model intelligible. Psychologists, meanwhile, may take the preference as a starting point, and want to pursue why the system works as it does. That research is unlikely to find holistic concepts like “preference” to be enduringly useful, except as an object of explanation. As analytic social science develops, points at which inadequate models yield empirical anomalies and failure may prompt efforts toward model of the actor that are less parsimonious and readily intelligible, but more internally realistic and behaviorally accurate. Appeals to more elaborate models of cognition in agent-based modeling, for example, may manifest just such a push for elaboration (Macy and Flache, this volume), as may all the efforts of the “heuristics and biases” work across cognitive psychology and behavioral economics (Kahneman and Tversky 2000; Camerer and Loewenstein 2004). Emphatically: intentional explanations are not cognitive explanations—and preference is not a cognitive concept—because, like rational choice explanations more generally, intentional explanations simplify precisely what the cognitive idiom complicates. Even so, and even as cognitive and neuropsychology continue to make “folk psychology” obsolete for its own purposes, we should not be surprised if the simple power of the concept of preferences sustains its vitality in analytic social science.

At the same time, analytic sociology may benefit from keeping the pragmatic justification for preferences in view for at least three reasons. First, we should not be surprised that interrogation can reveal all kinds of murkiness about exactly where preferences end and beliefs, capacities, emotions, etc., start, and we should not expect this murkiness to be subject to any tidy scientific resolution. Second, conceptualizing preferences as choice in a hypothetical “default” situation is a transparent fiction, and thereby highlights that real situations of choice are never abstract and never semantically empty. "Eliciting" preferences are not a matter of psychological excavation but of practical and social construction (Lichtenstein and Slovic 2006). Consequently, the idea of what to take as an individual's preference for any normative or scientific purpose rests on a defense of the procedure used that is not to be adjudicated by reference to "real" preference but other criteria. Third, we should not be especially troubled by failings of individuals to exhibit the coherence and consistency that application of the language of preferences would imply. The language of intentionality is the characterization of a system as a *unified subject*; whether natural persons act like unified subjects is fundamentally an empirical and contingent question. In other words, we should recognize that even as preference undergirds a powerful, ubiquitious idiom for understanding action, it is ultimately a logical concept loosely coupled to the actual acting organism, and so preference should be used when useful and unsentimentally discarded when some alternative idiom serves better instead. Likewise, in considering the rational (or reasonable) action theories versus more elaborate alternatives, we should keep in view that this is more a practical choice in terms of usefulness and parsimony for specific projects than any debate about what actor psychology is "really" like.

PREFERENCE AND PERSONALITY

In analytic social science, preferences are often assumed to be the same across persons. In Schelling's model, for instance, all actors were presumed to have the same preferences, and all variation in individual behavior in the model followed from variation in actors' immediate situations (the proportion of "like" neighbors). Especially when trying to explain variation in real behavior, however, social science regularly confronts actors behaving differently in *similar* situations. Variation in preferences serve as one candidate explanation.

Explaining choices with defensibly attributed preferences implies connecting the present choice to past choices or other evidence by virtue of abstract commonalities. One does not choose between an abstract apple and an abstract orange, but rather between *this* apple and *this* orange at *this* time under *these* conditions. Alternatives in real choices index abstract attributes that actors are asserted to have preferences regarding. When a student chooses to accept a scholarship at Northwestern instead of Yale, the explanation "I prefer Northwestern" suffices in one logical sense but is also obviously wanting for elaboration of *what it is* about Northwestern and Yale that caused the actor to choose one over the other.[2] Here, a useful distinction perhaps might be drawn between "preferences" and "tastes," with the latter being a more elemental subspecies of the former. That is, we might consider "taste" to index the attributes of alternatives leading one to be preferred to another, and "preference" to index the alternatives themselves (cf, Elster, this volume, on "substantive" versus "formal" preferences).

For decisions like what college to attend or where to live, alternatives obviously have many different attributes, and rarely is one alternative best in every respect. We thus need some way of talking about how tastes over different attributes of alternatives can explain the choice itself. In early utilitarianism, concrete preferences were posited as

[2] "Cause" here is used loosely. The statement does sustain a counterfactual, in that we are asserting that if specified attributes of Northwestern were different, the actor would perhaps prefer Yale instead. Yet, unlike usual understandings of "cause," attributes do not temporally precede what they are attributes of.

deriving from an ultimate preference to maximize happiness, and so preferences for concrete things in the world, could be interpreted via the expected consequences of their combination of attributes for happiness. Rational actor models have since come to discard the psychologism of "happiness" for the pure abstraction of "utility," which is nothing but what is maximized by a weighting of the preferences revealed by behavior (Friedman 1953; Schoemaker 1982). The "utility function" is the function by which different abstract "commodities" or "goods" that are manifested in different concrete alternatives are weighted. Following earlier arguments, the idea of an actual mathematical function precisely characterizing a long series of real-world individual behaviors becomes quickly quixotic, but, even then, utility functions provide a useful fiction for conceptualizing choices as the result of tradeoffs among abstract attributes. Even when sociologists eschew the formality of rational actor theories by proposing that actions are "reasonable" instead of "rational," a relaxed counterpart to the notion of a weighted function seems to be how they imagine different aspects of alternatives influencing ultimate choice (Boudon 1989; Hedström 2005).[3]

Exactly what are the abstract commodities indexed by a utility function is left vague. Even if we could articulate some specifiable commodity like "happiness" that served as the pertinent attribute of the utility function, it would remain to be explained what it is about choosing Northwestern over Yale yields the difference in expected "happiness." In explaining an action by reference to a preference as indicated by past choices, the explanation may be more satisfying the more extensive the past choices it invokes.

[3] The chief difference between "rational" versus "reasonable" action appears to be the uniqueness with which preferences can be said to determine actions.

Granted, some specific choices may be hard to understand beyond quite concrete tastes (as in a taste for vanilla), but others may manifest a more abstract pattern quite readily (as when a person makes a series of specific job decisions that manifest a taste for autonomy). If we want to make sense of heterogeneity in larger individual patterns of action, we might thus look for abstract tastes that are pertinent to choices across many situations.

Broadly relevant, abstract tastes that vary across persons may capture a large part of what we understand as *personality*.[4] Sociology's own literature on "social structure and personality"(House 1977) has shown only intermittent interest in the consequences of personality and has held a limited, unusual view of what personality is. Indeed, many sociologists persist in understanding their discipline as fundamentally about decrying constructs like personality as overrated for explaining action, and, accordingly, consequences of personality differences per se have been mostly left to psychologists. Yet perhaps personality psychology can provide analytic sociology with insight in *parameterizing the actor*, that is, help articulating what differences in abstract preferences may be especially useful in developing models for explaining behavioral differences.

Personality psychology has long sought a concise characterization of the dimensions of individual variation that have been salient and pervasive enough to become part of everyday language. Presently, the most highly regarded outcome of this effort is the Five Factor Model of personality (FFM) (John & Srivastava 1999; McCrae and Costa 2003). The FFM is not intended to "reduce" personality to five numbers, but it does propose that other personality constructs are associated with at least one of these five and that FFM provides a

[4] This implies that the basic problems of considering preferences as causes also apply to personality traits as causes.

useful simplification of the independent dimensions of individual variation. The dimensions of the FFM, sometimes called the Big Five, are Extraversion, Openness to Experience, Emotional Stability (also called Neuroticism), Agreeableness, and Conscientiousness.

Each dimension has facets. Inspecting a widely used FFM inventory (the NEO-PI-R) indicates that many facets resonate with abstract tastes suggested by sociological work on social order or status attainment (Hitlin and Piliavin 2004). Examples are facets that at least roughly index the value one attaches to the well-being of others, fulfilling commitments, obedience to traditional authorities, immediate gratification, social relations, achievement, dominance in situation, and being honest.

While many studies exist on the relationship between personality dimensions and individual outcomes of broader social science interest (see McCrae and Costa 2003: 216-233; Roberts et al. 2003: 580 for reviews and references), far less work has sought to integrate personality into the effort to understand connections between local interactions and larger social-level phenomena that characterizes the aspiration of analytic social science. Granted, in exercises focused on particular decisions—especially if they only are intended as illustrative using simulacra of actors anyway—considering preferences only in more superficial, specific terms may be more useful anyway. The abstract tastes indexed by personality traits may be most valuable when trying to understand patterns of substantively diverse choices or outcomes that are thought to be the outcome of a diversity of choices over the life course. The measurement of such tastes may be most useful in data collection efforts that are intended to serve many purposes and so may place a premium on extensively validated measures with possibly broad application.

At the same time, abstract tastes are only useful insofar as they actually cleave decision-making at its joints. Rational choice models of decision-making commonly invoke the concepts of *risk preference* and *time preference*. Although risk preference is often misunderstood by sociologists (see Freese and Montgomery 2007), the concept refers to the relative valuation of a gamble with an expected value of $*x* versus a certain payoff of $*y*. An actor who requires $x > y$ to choose the gamble is *risk-averse*, one who will take the gamble when $x < y$ is *risk-loving*, and one who is indifferent between the gamble and uncertain payoff when $x = y$ is *risk-neutral*. Risk preference can be posited as important to all kinds of different decisions, but some evidence suggests that individuals do not exhibit especially strong stability of risk preference across different domains (Berg, Dickhaut, and McCabe 2005; Freese and Montgomery 2007). Time preference refers to rate at which future rewards are discounted, as in the magnitude of the difference between *x* and *y* for which a respondent would be indifferent between receiving $*y* now and receiving $*x* in a year.[5] As with risk preferences, evidence seems to indicate that time preferences differ for different domains (Frederick, Loewenstein, and O'Donoghue 2002). Consequently, it remains unclear how useful it will prove to attempt to characterize individuals as having a general risk or time preference, even though the constructs hypothetically are applicable to a wide range of human decisions.

The relative neglect of the consequences of personality differences by sociologists may contribute to a broader underappreciation of how these consequences are determined by larger social processes. For example, analytic models demonstrate that the expected

[5] That discount rates are typically nonlinear can be quickly intuited from considering that the difference required for an actor to be indifferent between $*y* now versus $*x* in a year is perhaps much larger than between receiving $*y* in 1 year and $*x* in 2 years.

consequences of altruistic preferences vary depending on the number of fellow altruists among one's interactants. Tendencies toward impulsive or sensation-seeking behavior may have more negative outcomes for adolescents from disadvantaged backgrounds (Shanahan et al.forthcoming). Policies emphasizing lightly regulated individual choice may yield greater returns to conscientiousness, and periods of rapid technological advance may especially advantage those with high openness to experience (Freese and Rivas 2006). Additionally, although ample evidence exists for assortative mating and homophily of other familial and social ties by personality traits, little explicit consideration has been given to how this network clustering may amplify or dampen the effects of individual differences (a tendency for the gregarious to associate with one another, for instance, might imply a much greater effect on gregariousness on number of *second*-degree ties than if people formed such ties randomly, which might amplify whatever positive effects such ties provide). Studies of the import of individual differences warrant complementary consideration of what conditions make differences more or less important, and this kind of inquiry demands a more sustained sociological imagination than what studies by psychologists provide. In sum, analytic sociology's interest in modeling actor variation may find conceptualizing this variation as following from variation in preferences obviously appealing; reflective consideration of the root tastes involved may allow possibility for theoretical and empirical connection to well-studied concepts in personality psychology; and analytic sociology should consider as part of its project understanding how social dynamics can make preference variation more or less consequential for ultimate outcomes.

WHO HAS PREFERENCES?

In everyday life, we use intentionalist language promiscuously. We say "Sally wants to go to Europe" and also that "Wal-Mart wants to expand into Europe." The commonsense understanding is that we are using the language of intentionality literally when talking about "natural persons" and figuratively when talking about "collective subjects." An upshot of the first section, however, is that it may be more apt to consider that preferences are always being used figuratively, at least if by "literally" we mean to be referring to identifiable internal states of actors. Consequently, that preferences only really apply to individuals because preferences are properties of "a" brain cannot be defended. If preferences are characterizations of intentional systems, then the appropriateness of saying that a system is acting intentionally would seem to rest less on intuitions about the system's ontology and more on how well the application fits the system's behavior (Tollefson 2002; King, Felin, and Whetten 2007).

In some cases, attributions to collective subjects, like "Asian American adolescents value achievement and connectedness more than do Caucasian Americans" (Hitlin and Piliavin 2004: 369), are based just on a simple aggregation of individual behavior; there is no pretense of a system whose action is being explained. Why this tendency among individual persons exists can be taken as a matter of sociological explanation, perhaps drawing on some specific experience, identity, or goal held in common. Other collective subjects, however, are understood as exhibiting coordination among constituent individuals. The activities of workers on an assembly line can be made collectively intelligible as "building a car." Coleman (1990) uses "corporate actors" to refer to constructed relations of multiple natural persons that may be otherwise characterized as “actors,” and the prototype for such actors might be the corporation, at least as classically envisaged as a kind of intelligent, purposive social machinery.

Talking about organizations as acting according to preferences is a way of making sense of organizational behavior while black-boxing the internal dynamics that result in their behavior as intentional systems. In amny cases, that black-boxing may be undesirable to social science, as understanding the internal process that leads to an organization doing one thing rather than another may be precisely the object of a sociological inquiry. Nonetheless, in principle, interpreting an organization’s activities as those of a unified subject is sustainable so long as those activities collectively support the attribution of either stable or intelligibly changing beliefs and preferences. Indeed, by this standard, organizations may exhibit action more coherent and consistent with attributed preferences than do natural actors. Applications of rational choice theories, notably, often characterize actions of firms and political parties much more effectively than that of natural persons (Satz and Ferejohn 1994; Clark 1997), for several reasons. First, organizations are typically brought into being for accomplishing a specific purpose, like enriching their principals, which then might serve as dominating the preferences of the actor. Second, organizations may exist in environments that strongly discipline actions that stray from purposes necessary for continued existence of the organization (e.g., earning a profit). By contrast, natural people can tolerate a very broad range of nonruinous preferences in typical environments. Third, while we may have the experience of there being some place in our brains where “everything comes together” for deliberative decision-making, strong indications from cognitive science are that decision-making is much more decentralized (Dennett 1991). Organizations, on the other hand, commonly have a centralized decision-making unit that can explicitly direct action, monitor the effectiveness of different coordinated parts in implementing actions, and amend its structure to better realize actions. Fourth, in making decisions and coordinating actions, organizations may develop more externalized, coherent, and binding statements of identity that become a guide for subsequent action (King, Felin, and Whetten 2007). Finally, in bringing together multiple decision makers and specialists, organizations may be less prone to simple "performance errors" in trying to figure out the best course of action.

Nonetheless, organizations often fail to exhibit the stability of purpose that fit the characterization of them as a unified subject. From the outside, such failures may look like instability of purpose over time, a lack of coordination among parts, ineffective pursuit of purposes, failure to engage seemingly obvious opportunities, or incoherence of various parts of the organization. Explaining the failure of an organization to act like a unified subject will often then proceed by opening up the organization and looking at the actions of its agents from the standpoint of considering those members as intentional systems acting in structural relationships with one another. Of course, when organizations *do* act like unified subjects, exactly how this is accomplished *also is a matter for inquiry*.

When talking about the preferences of natural persons, we considered how intentionality provided a partition and interface between the labors of analytic social scientists and psychologists. With respect to the preferences of organizations, sociologists are often interested precisely in understanding why the organization acts as it does. More quickly than with natural persons, social scientists perceive intentionalist language as applied to organizations as inadequate and look for explanatory tools that talk about how structure and process within the organization produce behaviors only intermittently characterized as that of a unified subject. However, such a tendency need not reflect any actual fact of nature, but instead just the questions that the social scientist is asking and the availability of information on internal processes. The membership of an organization, the structure of their relations, and the content of their interactions are components of organizational functioning that do not take place in the heads of individual members, and these provide core materials for social scientists interesting in the internal dynamics of organizational behavior.

At the same time, our promiscuity in using intentionalist language does not just extend to units of analysis larger than ourselves. We sometimes say things like "part of me wants to quit my job." People regularly characterize their own deliberation as manifesting various kinds of internal tension, inner dialogue and of-two-mindedness. The notion of people possessing multiple intentionalities has been the anchor of one prominent line of sociological social psychology (Mead 1934), and discussions of "the unconscious" by both psychological and sociological theories has often proceeded as though it were a separate intentional system inside us. When talking about conflict within organizations, we can readily understand outcomes as the product of a negotiation between two or more opposing subjects in conflict, and our sense of the reality of such conflict is enhanced by the sides being manifested in real persons with "real" intentions. When talking about inner conflict, it seems unclear what comprise these sparring ghosts in the machine.[6] Perhaps this is all just misleading introspection, or we may actually be able to look at moment-to-

[6] Speaking of ghosts in the machine, Coleman (1990: 526) speculates about modeling the lack of the unity of purpose in natural persons as multiple intentionalities based on internalized models of primary socialization agents, akin to having the wishes of your mother and first-grade teacher fighting inside your head.

moment behaviors in our lives and find it hard to reckon that we have acted with an overarching unity of purpose, as opposed to appearing to dither over vacillating dominant ideas of what we want.

Some theorists have suggested that we might usefully think about natural persons as a squabbling "parliament" of interests, which can be thought of as a narrow set of preferences (Ainslie 2001; see also essays in Elster 1985).[7] Dynamics among these interests may then provide a better characterization of behavior than any alternative wed to our being a truly unified subject, as opposed to just sometimes looking like one. Useful here also may be that culture provides schema for the intentionality implied by different identities, and one can have preferences about manifesting these identities quite apart from the preferences that comprise them ("I don't enjoy peer-reviewing papers, but it's part of being an active member of the research community"). Identity theories often conceive of the self as internalized hierarchy of self-categorizations (Stryker 1980; Stets and Burke 2000) that bears no small affinity to a preference ordering, and each of these categorizations carries socially-acquired understandings of the kind of behavior an intentional system appropriate to that categorization would manifest. (as in understanding how the *performance* of an ideal member of this identity would be carried out, a la Goffman [1959]). In this respect, instability as a unified subject would follow from the uncertain resolution of conflicts, not just about what we want but about what kind of person we want to be. Indeed, Winship (2006) offers the metaphor of puzzle-solving to describe the actual muddling work of policy design—an effort to engineer solutions that satisfy competing

[7] While perhaps more evocative of simply scattered rather than multiple but discrete intentionalities, Gould (2003) offers some especially intriguing speculations about using network ideas to talk about the coherence of temporally successive actions.

preferences as well as possible—and perhaps the same kind of puzzle-solving is characteristic of the individual engineering of a busy life. In sum, sociologists should find no discomfort in using preferences to refer to collective subjects, and sociologists interested in the complexity of selves may profit from trying to consider more systematically how individuals may manifest multiple intentionalities.

PREFERENCE FORMATION AND CHANGE

In analytic social science, preferences are commonly posited not just to be the same across persons but also unchanging over time. The immutability of preferences has received especially vigorous defense in economics, as the ability of models to generate unique, precise predictions is much complicated by preferences being endogenous to unfolding events. Stigler and Becker (1977: 76) provide an especially rousing description of the working premise that "tastes neither change capriciously nor differ importantly between people. On this interpretation one does not argue over tastes for the same reason that one does not argue over the Rocky Mountains—both are there, will be there next year, too, and are the same to all men." This statement might seem remarkable for how it is flatly contradicted by even casual observation—the very notions of personality and socialization, for instance, depend on preferences differing nontrivially across people and over time. For that matter, discussions of emotion and action often afford interpretation as transient changes in preferences, as in the idea of rage involving an alteration in the preference for immediate vs. delayed outcomes and in preferences regarding the well-being of others (Elster, this volume).

Stigler and Becker's ultimate position is far less radical than this pronouncement. As discussed earlier, the tangible objects of choice are taken only to be *indirect* means of obtaining the commodities that are the *real* objects of preference in an analysis. Stigler and Becker discuss several ways that experiences might affect various stocks of psychological "capital," which in turn modify the extent to which tangible choices provide the commodities sought. So, specific music preferences change but the abstract preference for activities yielding "appreciation" remains the same; the specific preference for, e.g., "Blue Danube" over "Well-Tempered Clavier" changes over time because experience allows the actor to appreciate "Well-Tempered Clavier" more than before.[8] These abstract preferences are what I have been here calling "tastes." By emphasizing abstract tastes for commodities, the model allows analysis to proceed with the assumption these tastes are unchanged, but at the same time offering what can be interpreted as a simple model of learning or socialization. Indeed, the model is ultimately quite consonant with the familiar sociological emphasis on past experience as explaining psychological variation across persons. Far more incompatible with the approach is behavioral genetics, with its body of evidence that as close to half of intrapersonal variation in personality traits may be accountable by genetic variation between persons (Loehlin 1992). Ironically, then, even as many sociologists express antipathy toward both neoclassical economics and behavioral genetics, the greater the relevance of genetics for understanding behavioral variation, the greater the challenge to orthodox economic assumptions about preferences.

Sociology is distinct from both behavioral genetics and economics in the extent to which the malleability of preferences over the life course has played a central role in

[8] Example from Bourdieu (1984).

theorizing. For instance, crucial to arguments about the importance of norms is that norms influence behavior not just by affecting incentives (i.e., via the threat of sanctions), but that they are *internalized*, experienced by the person as reflecting right and wrong, good and bad (Hechter, this volume). *Association* models provide the dominant means of conceptualizing how preferences change by social scientists. In the prototypic example, actors are (repeatedly) exposed to some X and also to some Y for which they have a strong preference or otherwise positive affective reaction. As a result, they come to have a preference for X even in the absence of Y. This Y is sometimes thought of as some kind of more fundamental preference ("need" or "drive"). As described by Coleman (1990: 516):

> [T]he course of events and action creates over time a whole superstructure of interests [i.e., preferences] that were originally pursued as paths toward the satisfaction of more central needs. These come to be autonomous interests insofar as they satisfy, over some period of time, the more central needs.

Structural-functionalism's faith in the great power of relatively simple conditioning led to Wrong's (1961) complaint of an "oversocialized" conceptualization of actors and Garfinkel's (1967) complaint of actors being taken for "cultural dopes."

While structural-functionalism has fallen somewhere beyond mere disrepute, sociologists still grant "socialization agents" considerable capacity to purposively influence the preferences of young actors (Hitlin and Piliavin 2004). Parents have preferences regarding the preferences their children develop, and variation across parents may dampen intergenerational mobility. Kohn (1969) presents evidence that upper-class parents emphasize teaching their children self-direction while lower-class parents emphasize obedience; these preferences are speculated to be useful for subsequent status attainment and thereby set the stage for reproduction of the same divergence of socialization practices and attainment in the next generation. Status attainment theory posits that parental encouragement influences aspirations, which in turn influences choices about investing in schooling (Sewell and Hauser 1975). Bourdieu's theory of investments by parents in the cultural capital of children highlights the possible contribution of aesthetic preferences to subsequent status attainment (Bourdieu and Passeron 1990).

Bourdieusian theory also reflects a longtime recognition by sociologists that preferences are used by others to make broader inferences about an actor. As Bourdieu (1984: 6) famously put it, "Taste… classifies the classifier." Individuals make favorable attributions toward others perceived as having either preferences associated with high status or preferences like their own. Indeed, a standard protocol for ingroup bias experiments in social psychology involves asking respondents to express a preference between paintings by Kandinsky and Klee, and subjects exhibit favoritism toward others that they are led to believe expressed the same preference (Tajfel et al. 1971).[9] The social significance of preferences provides incentives for actors to lie about their preferences, or to selectively and conspicuously advertise those preferences that are associated with esteem or similarity to others.[10] Actors also have incentive to invest in purposive

[9] Valuing the well-being of another person more after forming social ties with them is a species of preference change.

[10] Elster's writings on rationality proceed by copious, conspicuous reference to high culture sources for examples about human life (e.g., Elster 2000, 2007); an interesting thought experiment is to imagine how the same logical arguments would read differently if examples from television or other popular culture sources was used instead.

cultivation of preferences consonant with a preferred identity or set of social affiliations.[11] In other words, purposive manipulation of preferences is not just for socialization agents, but can also be self-directed, as in someone who takes a course about wines out of a desire to appear more sophisticated or to be able to banter knowledgably with vino-erudite consociates. That said, it is unclear how much of the influence of identity on preference development can be characterized as something like "investment" in "appreciation," as opposed to having aesthetic preferences being more fundamentally tied to the idea of those perceived as like or unlike them. The distaste that contemporary educated adults have toward names for newborns like "Crystal" and "Jimmie," for instance, seems to derive *fundamentally* from the déclassé connotations these names have acquired (Lieberson 2000).

Preferences may provide the most convincing signals for identity precisely when they are, in a sense, "unnatural." Preferences that require motivation or work to cultivate, and perhaps even are hard to cultivate at all without an extensive background base of knowledge and experience, can service cultural boundaries well (i.e., preferences provide a kind of signal, see Podolny and Lynn, this volume).[12] This observation can be juxtaposed with the point that something like a default "human nature" is implicit in many

[11] Entrepreneurs have strong incentives to devise and engineer associations among individuals between proprietary goods and preferred identities and affiliations. How goods come to have the implications for identity and affiliation that they do is itself an important topic for the sociology of consumption.

[12] If a preference that distinguishes a group also confers status within it, this might promote runaway processes that culminate in claims of finding beauty in stimuli so extreme that an outside observer might wonder whether everyone who professes to the preference is just pretending. In this respect, aesthetic preferences within subcultures may manifest phenomena similar to sectarian splits and zealotry in religious belief.

considerations of preference change.[13] Roughly, this natural state is something more selfish, hedonic, and immediate-reward oriented than the mature actor whose preferences have been shaped by various socialization and other civilizing processes of contemporary society. "Incomplete" socialization was one time prominent in sociological explanations of deviant behavior (e.g., Dubin 1959), and self-control based theories posit that preferences for delayed reward are something that institutions cultivate in actors. Bowles and Gintis's (1976, 2000) account of the educational system may be thought of as attempted preference modification away from a state of nature and toward preferences more suitable for a capitalist economy. In short, "unnatural" preferences may be rewarded in social systems—perhaps partly because they are unnatural—and these rewards may prompt investment in their cultivation, making preferences endogenous to the demands of society. Sociobiologically-minded scholars have frequently contended that "human nature" serves to keep societal variation "on a leash" (Lumsden and Wilson 1981: 13; Udry 2000). An important counterpoint is that social competition may create incentives for cultivating and otherwise engineering the self toward extremes.

As suggested earlier, identities may prompt individuals to engage in actions indicative of a preference for a different set of preferences than what their behaviors to that point imply. If we use *goals* in a fairly restrictive sense to refer to non-immediate, articulable attainments that actor can express (possibly sporadic and floundering) commitment to pursue, then the same can be said of goals as well. An actor having

[13] "Human nature" is most commonly invoked for explaining preferences held in common among persons, such as the desire for status or social affiliations (Turner 1987). Evolutionary historical explanations are commonly provided for why we have such preferences, although the fundamentally speculative character of many such accounts is well-known.

preferences about preferences is a tricky notion—after all, if we really wanted not to want something, haven't we already succeeded?—but goals and identities carry implications for preferences and can be objects of preference in their own right.[14] As such, these implications can stand in conflict with one another and with tastes for abstract commodities like "hedonic pleasure." Internal processes by which these conflicts are handled may lead to introspective experiences and behaviors that undermine easy characterization as a unified subject in folk psychological terms. Such internal processes also have external counterparts in efforts of others to commit actors to different goals or identities and to shape belief about what different goals and identities oblige. Sociology's openness to preference change allows for a much more nuanced consideration of *manipulation* of actors—whether by others or by actors themselves—than what orthodox rational actor models allow.

REFERENCES

Ainslie, George. 2001. Breakdown of Will. Cambridge: Cambridge University Press.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 2005. "Risk preference instability across institutions: a dilemma." Proceedings of the National Academy of Sciences. 102: 4209-4214.

Boudon, Raymond. 1989. "Subjective Rationality and the Explanation of Social Behavior." *Rationality and Society* 1:173-196.

[14] Indeed, *setting a goal* implies a desire to produce actions that one would not otherwise produce, and then consistency with goals becomes another attribute to influence choices.

Boudon, Raymond. 1998. "Social Mechanisms without Black Boxes." Pp. 172-203 in *Social Mechanisms*, edited by Peter Hedström and Richard Swedberg. Cambridge: Cambridge University Press.

Bourdieu, Pierre and Jean-Claude Passeron. 1990. *Reproduction in Education, Society and Culture*. Newbury Park, CA: Sage.

Bowles, Samuel and Herbert Gintis. 1976. *Schooling in Capitalist America*. New York: Basic Books.

Bowles, Samuel and Herbert Gintis. 2000. "Does schooling raise earnings by making people smarter?" in *Meritocracy and Economic Inequality*, edited by K. Arrow, S. Bowles, and S. Durlauf. Princeton, NJ: Princeton University Press.

Camerer, Colin F. and George Loewenstein. 2004. "Behavioral Economics: Past, Present, and Future." in *Advances in Behavioral Economics*, edited by C. F. Camerer, G. Loewenstein, and M. Rabin. Princeton: Princeton University Press.

Camerer, Colin F., George Loewenstein, and Drazen Prelec. 2005. "Neuroeconomics: how neuroscience can inform economics." Journal of Economic Literature 43:9-64.

Clark, Andy. 1997. Being There: Putting Mind, Body, and World Together Again. Cambridge, MA: MIT Press.

Coleman, James S. 1990. Foundations of Social Theory. Cambridge, MA: Harvard University Press.

Costa, Paul T. and Robert R. McCrae. 1992. *NEO-PI-R Professional Manual*. Lutz, FL: Psychological Assessment Resources.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dubin, Robert. 1959. "Deviant Behavior and Social Structure: Continuities in Social Theory." American Sociological Review 24:147-164.

Elster, Jon, ed. 1985. *The multiple self.* Cambridge: Cambridge University Press.

Elster, Jon. 2000. Ulysses Unbound: Studies in rationality, precommittment, and constraints. Cambridge, UK: Cambridge University Press.

Elster, Jon. 2007. "Explaining Social Behavior: More Nuts and Bolts for the Social Sciences." Cambridge, UK Cambridge University Press.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." Journal of Economic Literature 40:351-401.

Freese, Jeremy. Forthcoming. Genetics and the social science explanation of individual outcomes. *American Journal of Sociology*.

Freese, Jeremy and Salvador Rivas. 2006. "Cognition, Personality, and the Sociology of Response to Social Change: The Case of Internet Adoption." *CDE Working Paper, No. 2006-07*.

Friedman, Milton. 1953. "The Methodology of Positive Economics." Pp. 3-43 in *Essays in Positive Economics*: University of Chicago Press.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, N.J.: Prentice-Hall.

Gieryn, Thomas F. 1999. *Cultural Boundaries of Science: Credibility on the Line*. Chicago: University of Chicago Press.

Gintis, Herbert. 2007. "A Framework for the Unification of the Behavioral Sciences." Behavioral and Brain Sciences 30:1-61.

Goffman, Erving. 1959. The Presentation of Self in Everyday Life. Garden City NY: Doubleday.

Gould, Roger V. 2003. Collison of wills: How ambiguity about social rank breeds conflict. Chicago: University of Chicago Press.

Hedström, Peter. 2005. Dissecting the Social: On the Principles of Analytical Sociology. Cambridge: Cambridge University Press.

House, James S. 1977. "The Three Faces on Social Psychology." *Sociometry* 40:161-177.

Jensen, Michael C. and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure." Journal of Financial Economics 3:305-360.

John, Oliver P. and Sanjay Srivastava. 1999. "The big five trait taxonomy: History, measurement, and theoretical perspectives." Pp. 102-138 in *Handbook of Personality Theory and Research*, edited by L. A. Pervin. New York: Guilford.

Kahneman, Daniel and Amos Tversky. 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

King, Brayden, Teppo Felin, and David A. Whetten. 2007. "What does it mean to act?: The theoretical foundations of the organization as a rational actor." Working Paper, Brigham Young University.

Kohn, Melvin L. 1969. *Class and conformity: a study in values*. Homewood, IL: Dorsey Press.

Lichtenstein, Sarah and Paul Slovic. 2006. "The construction of preference: an overview." Pp. 1-40 in *The Construction of Preference*, edited by S. Lichtenstein and P. Slovic. Cambridge, UK: Cambridge University Press.

Lieberson, Stanley. 2000. *A Matter of Taste*. New Haven, CT: Yale University Press.

Loehlin, John C. 1992. *Genes and environment in personality development*. Newbury Park, CA: Sage.

Lumsden, Charles J. and Edward O. Wilson. 1981. *Genes, Mind, and Culture: The Coevolutionary Process*. Cambridge, MA: Harvard University Press.

Marwell, Gerald and Pamela E. Oliver. 1993. The critical mass in collective action: a micro-social theory Cambridge: Cambridge University Press.

McCrae, Robert R. and Paul T. Costa. 2003. *Personality in Adulthood: A Five-Factor Perspective, Second Edition*. New York: Guilford Press.

Mead, George Herbert. 1934. Mind, Self, and Society from the Standpoint of a Social Behaviorist. Chicago: University of Chicago Press.

Roberts, Brent W., Richard W. Robins, Kali H. Trzesniewski, and Avshalom Caspi. 2003. "Personality Trait Development in Adulthood." Pp. 579-595 in *Handbook of the Life Course*, edited by J. T. Mortimer and M. J. Shanahan. New York: Kluwer.

Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge, MA: MIT Press.

Samuelson, Paul A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Satz, Debra and John Ferejohn. 1994. "Rational Choice and Social Theory." The Journal of Philosophy 91:71-87.

Schelling, Thomas C. 1978. Micromotives and Macrobehavior. New York: W. W. Norton.

Schoemaker, Paul J. H. 1982. "The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations." *Journal of Economic Literature* 20:529-563.

Sewell, William H. and Robert M. Hauser. 1975. *Education, Occupation, and Earnings: Achievement in the Early Career*. New York: Academic Press.

Shanahan, Michael J., Stephen Vaisey, Lance D. Erickson, and Andrew Smolen. Forthcoming. Environmental contingencies and genetic propensities: Social capital, educational continuation, and a dopamine receptor polymorphism. *American Journal of Sociology*.

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

Stryker, Sheldon. 1980. *Symbolic interactionism: a social structural version*. Menlo Park, CA: Benjamin Cummings.

Tajfel, H., M. G. Billig, R. P. Bundy, and C. Flament. 1971. Social categorization and intergroup behavior. *European Journal of Social Psychology* 56: 364-373.

Tollefson, Deborah. 2002. "Organizations as true believers." *Journal of Social Philosophy* 33:395-410.

Turner, Jonathan H. 1987. "Toward a sociological theory of motivation." American Sociological Review 52:15-27.

Udry, J. Richard. 2000. "Biological Limits of Gender Construction." *American Sociological Review* 65:443-457.

Winship, Christopher. 2006. "Policy Analysis as Puzzle Solving." Pp. 109-123 in The Oxford Handbook of Public Policy, edited by M. Moran, M. Rein, and R. E. Goodin. Oxford, UK: Oxford University Press.

Wrong, Dennis H. 1961. "The Oversocialized Conception of Man in Modern Sociology." *American Sociological Review* 26:183-193.