Economics and Philosophy, **20** (2004) 375–416

REVIEWS

DOI: 10.1017/S0266267104210318

The Scientific Study of Society, by Max Steuer. Kluwer Academic Publishers, 2003, xiii + 464 pages.

Max Steuer's readable book offers both an introduction to contemporary work in social science and also a defense of some general views about the nature of this kind of inquiry. Practicing social scientists will likely warm to its instinctive sympathy for their work. What of philosophers? Although both the author and Ken Binmore in the foreword are eager to deny that this book is an exercise in philosophy, its central claims – that a scientific study of society is possible and that its method is distinct from other ways of producing social knowledge – express meta-propositions about social science. What is distinct about Steuer's approach is his conviction that these questions are best addressed not through abstract argument but rather by carefully examining what social scientists actually do. In this spirit, while chapters in the beginning and at the end of the book contain his general, or philosophical, discussion, at the heart of Steuer's inquiry are six central chapters comprising long and painstaking reports of actual research. By the author's own admission, the philosophical discussions at either end of the book are of a rather informal nature and do not seek to engage explicitly with the philosophical literature. Rather, the rhetorical strategy is one of argument by illustration. Does it succeed?

The arguments presented in the early chapters are typical of a broadly naturalistic view of social science. Thus social science's goals are taken to be similar in kind to those of natural science, and its relatively bad empirical record to be explained by a number of practical disadvantages it faces. One is that the phenomena studied by social science are subject to change at a much greater rate. Physical and biological phenomena also change, but many of their underlying principles are both quite stable and also directly relevant to explanation and intervention. In the social world, the underlying principles (for example, self-interested behavior in the

economic sphere) may be relatively stable too. However, the fast-changing superficial features, such as credit cards, television and computers, are often the ones most in need of explanation as well as being powerful agents of change in their own right. This alone makes it difficult to isolate categories and identify reliable causal relationships. A second disadvantage is that controlled experiments, often taken to be the gold standard of causal inference, are much less available to social scientists. Finally, because the questions investigated by social scientists bear a closer relationship to our ethical and political allegiances, the threat of bias and loss of objectivity is greater. For their part, the two chapters at the end of the book comprise a self-consciously cautious attempt to draw general conclusions, covering four broad areas. First, Steuer seeks to provide a general characterization of each discipline on the basis of their principal foci of investigation and methodologies; second, to analyze and defend current disciplinary divisions; third, to pass a verdict on the current state of social science; and fourth, to argue for greater use of social science in public policy.

Being aimed more at lay readers, on their own these chapters will probably not convince those philosophers who are unsympathetic. Therefore much weight falls on the six illustrative chapters at the heart of the book. Steuer's method is to trawl systematically through ten years of top journals in each of what he considers the five major social sciences, namely anthropology, economics, political science, social psychology, and sociology. He reports papers in those journals bearing on six selected topics: crime, migration, family, money, housing, and religion. In effect, the reports are then expected to speak largely for themselves: to show that social science is a feasible project, that it is better than the alternatives, and that those in positions of power should take its claims seriously. The hard work, ingenuity and sheer intellect that went into many of these studies is indeed impressive, and well demonstrated by such extended illustration. The topics are chosen in such a way as to maximize coverage of similar issues by different disciplines from different perspectives. In reviewing the journal articles, Steuer reports the problems addressed by researchers, the conclusions arrived at and the methods employed.

Objections could be raised against the sampling procedure. Excluded from the survey are books (as opposed to journal articles), work not in English, papers in minor and specialist journals, work appearing directly on the internet, plus, of course, work on topics other than the six of interest. However, at least in the case of economics, books and mainstream work not in English tend to be similar in style to the journal articles surveyed, and minor journals and unrefereed internet material do not represent the mainstream in the first place. It is unclear whether the picture is quite so sanguine in the other social sciences, especially with respect to books and to work not in English. Nevertheless, we agree with Steuer that overall,

at least for the purpose of comparing it with alternative approaches (see below), his method illustrates well enough social science as it is actually practiced.

A major virtue of these central chapters is their demonstration of the diversity of interests, methods and epistemic categories that social research produces. Firstly, different disciplines are interested in different aspects of the six broad topics. For example, when it comes to crime social psychologists are preoccupied with rape, economists more with white-collar crime. Secondly, what is taken to constitute appropriate data varies substantially. For instance, interviews with subjects are an essential part of hypothesis-building for social psychologists and anthropologists. But they matter little to economists who, rather than gauge motivations empirically, instead just make assumptions about them that they hope will be widely applicable. Thirdly, attitudes toward the appropriate methods for testing causal claims are also shown to vary widely: in economics (or at least some subfields of it) it is commonplace to accept or reject a causal hypothesis purely on the basis of its demonstrability in a mathematical model, but in sociology statistical tools are often used instead. Finally, Steuer makes some fascinating observations about the variety of epistemic categories at play in social science. In the chapter on family, he recounts how facts can be variously "direct" (uncontroversially verifiable), "contextual" (invoking broader social tendencies), "compiled" (statistically aggregated), "stylized" (challengeable interpretations) and "high order" (claims about relations of facts).

So how successful for his larger purposes is Steuer's strategy of argument by illustration? To answer that, it is necessary to be clear on who his targets are. Besides aiming to introduce the field to lay readers, another explicit motive is to contrast social science against the work of populist "frauds and impostors" (54) who "do something unscientific and pretend it is social science" (409). A second target are "people in universities who are antagonistic to science in any form" (17), and who doubt that society can be studied scientifically (409). A third complaint (chapter 12) is against what Steuer sees as the active ignorance of social science on the part of laypeople and policymakers. We may label these three targets informally as: "quacks," unfriendly academics, and an ignorant public. We judge that the book succeeds against the first of these, not against the second, and only partially against the third.

Start with the first category. Steuer takes art, history and philosophy to be valid alternatives to social science, because these endeavors assume goals explicitly distinct from scientific explanation without "pretending to be social science" (54). The invalid alternatives come largely from outside universities. Steuer divides these activities into "social revelation," "social criticism" and "social poetry" (55–62). The first includes popular attempts to explain all social phenomena by reference to some one

overarching insight, such as network, risk, consumerism and so on. These characteristically do not draw on or respond to the large body of relevant empirical or theoretical work that exists in mainstream social science, instead relying on grand revelations aimed at accounting for social reality as a whole. Social critics, for their part, denounce problems in society through films, books, TV programs and the like. Their targets include the hypocrisies of suburban life, the media, global capitalism, etc. Finally in Steuer's taxonomy, social poets aim at a primarily emotional impact by creating "penny-dropping" artistic experiences about life in the modern world.

Many populist examples of social analysis are indeed made to seem simplistic and ignorant merely by Steuer's prolonged recounting of mainstream social science in action. To the extent that they want to claim the authority of social science, we thus judge the book to be an effective strike against them.

Turn now to the second category, and the more sophisticated threat represented by unfriendly academics, presumably those found in cultural theory, literary studies or other fields inspired broadly by twentieth-century Continental philosophy. A strand of these movements is taken to denounce rationality and the scientific outlook as a whole, or else to reinvent the purposes of social science completely. These opponents unfortunately are not so easily dealt with. Perhaps Steuer is reacting against what can sometimes seem a willful ignorance of social science on their part. Nonetheless many of their arguments require responses beyond simple illustration of existing social scientific work, as the latter on its own cannot make the case Steuer wants it to without better philosophical packaging.

One example of this is the claim in chapter 2 that science is characterized as a collective enterprise of building a "structure." The building blocks of this structure are explanations of phenomena, which in turn rest on other explanations. These pieces need not fit together neatly, and connections between different explanations can be suspected, established or wholly non-existent. Nor does the structure need to be understood hierarchically. Rather, Steuer argues, the important features are that each individual scientist is working *within* and *in response to* the host of explanations put forward before her and that her claims in turn are subject to peer review. But we are skeptical whether this criterion alone is enough to demarcate social science from what Steuer calls "pretend social science" (424). Perhaps no demarcation can be expected to work smoothly everywhere, but arguably this one fails even at the initial stage. Postmodernism, poststructuralism, cultural studies and other approaches Steuer wishes to denounce all inspire work that could be described as providing understanding via building their own structures of connected explanations. Granted, these explanations appeal to factors

very different from those one finds in mainstream social science, but Steuer does not tell us why this difference matters and neither does his proposed criterion.

Turning to the third category, outside academia public awareness of social science is poor. Indeed, unlike natural science it is systematically ignored by the very people who could make best use of its findings, i.e. policymakers. Steuer concludes in the last chapter that this active ignorance is scandalous. A lot of current social science clearly aims at producing knowledge that in some way can be relevant for public policy. Yet, with the exception of economics, whose high prominence in government policy Steuer credits to the work of J. M. Keynes, social scientific research rarely figures in relevant public debates. Typically, even when facing questions that lie in the direct areas of competence of sociologists or political scientists, politicians and ordinary folk alike address them instead using just common sense or ideology. When discussing an issue of natural science, it is normal to defer to the relevant experts. By contrast, politicians – and for that matter lawyers, journalists, actors and sports personalities – are all too often happy to take their own unvarnished opinions as the first and last word on any matter of social science.

This point is, we think, well taken. Indeed the barb might be extended further to the many *academics* who, when straying beyond their own fields of competence, frequently end up in the realm of some social science. Nevertheless we judge the book only partially successful against its third target because its argument here relies crucially on the claim that social science is indeed useful, at least potentially, for policy. True enough, many of Steuer's examples do show that it illuminates particular issues well beyond what is possible from the armchair or by reading the newspapers, but still the claim to policy relevance often does not hold up. Steuer, as an experienced practitioner, is well aware of this problem and gives two responses (415): first, scientifically informed uncertainty is itself a valuable piece of knowledge, better than any other ground for policy choice. Second, greater attention to social science on the part of governments will itself tend to improve matters. Economics, Steuer claims, is a case in point. Since the mid-twentieth century, in the UK and US economists have played a prominent role in advising government on questions of economic policy. This involvement, Steuer thinks, has by itself spawned a wide range of applied work in many areas of the discipline, and he hopes that the same support of applied research can be given to other social sciences. These responses are interesting, although perhaps made rather quickly.

A more fundamental difficulty here though is not, we think, addressed adequately: just what methodologies will generate potentially applicable knowledge? Social science as illustrated in the book's six central chapters does not face the ignorant enemy as a united front. Rather, as noted earlier,

the picture resembles instead an extremely diverse mixture of projects whose aims, standards and methodologies appear to bear little relation to one other. As a result, it is not left sufficiently clear what it is about social science that differentiates it from inferior alternatives. By itself diversity of methods and standards is not a vice, but it becomes so when it impedes our ability to detect and to integrate policy-relevant information. Perhaps in order to be heard by the public some consolidation of standards is required, or else an explicit articulation of just how the different methods each further a common goal. Steuer claims that for all social sciences the objective is the same (366), by which he must mean that all strive to provide the same kind of understanding. But in order to make the case for social science as a whole, this crucial point needs to be fleshed out. In particular, what is required is a more critical and rigorous analysis of when and why social science does and does not succeed.

A start would be to modify Steuer's conception of the goals of social science, since understanding via connecting explanations is far too vague a criterion. We mention one possible way of doing this here. Following J. S. Mill, one could focus on tendencies, i.e. the concrete causal forces that operate in the social world and that combine to make up social phenomena as we observe them. On this view, social science is the study of the identity and nature of these tendencies and the rules for their composition. Ideally, such knowledge then licenses successful policy intervention. Adopting such a picture would force social scientists to make more explicit how their different methods bear on this goal. Economists, for instance, regularly postulate tendencies in models but less often study how these tendencies are instantiated in the complex environments of the real world. Anthropologists, on the other hand, pay much attention to formulating the right categories for analyzing particular communities, but less to these categories' causal connections. Both can be seen as ways of approaching the study of tendencies.

The point is that a better conception of methodological goals may both improve social science and also make clearer the inadequacies of its rivals. To illustrate, a familiar complaint against rational choice theory is its practice of deriving causal relations from extremely idealized models and then claiming that *ceteris paribus* they obtain in reality. Although Steuer acknowledges the issue, he calls it only a "pretend problem" (42) by comparison to that of convincing policymakers to take social science seriously. We disagree: To be taken seriously, rational choice theorists must show how tendencies in models relate to tendencies in the world. Only then can a persuasive case be made for heeding their advice over that of others.

Despite these problems, we hope that Steuer's book will mark an important beginning. Given its lack of engagement with the philosophical literature it would be easy for philosophers to dismiss it, but we think this

would be to ignore the fact that its virtues – principally, a wide knowledge of and hence feel for social science as it is actually practiced – are the very ones most lacking in that literature. An informed discussion of the nature of social science, what can be expected of it, how it can be improved and how to bring it to bear on policymaking, is badly needed. In particular, rather than yet another recounting of general metaphysical obstacles like multiple realizability, much more attention should be given instead to why social science sometimes *does* succeed and to the methodological problems that are actually pertinent 'on the ground'. Steuer breaks the silence and one can only hope that the discussion will continue.

Anna Alexandrova

University of California, San Diego

Robert Northcott

London School of Economics

DOI: 10.1017/S0266267104220314

Rationality and Freedom, by Amartya Sen. Harvard University Press 2003.

In *Rationality and Freedom*, Amartya Sen invites readers to acknowledge that human freedom is a key concept in philosophy, economics, and social science – not to mention the highest value in contemporary free-market democracies – and simultaneously to realize that it remains an elusive concept (583–585). Those familiar with Sen's bibliography will be grateful to have a selection of his writings tailored to discuss his core interests in social choice, justice, and rationality. His text is a compendium of essays, written between 1983 and the present ;[1] readers interested in social choice or the foundations of decision theory will benefit from having this collection at their finger tips both for reference and as a statement of Sen's overarching philosophical position on rationality.

This review provides a synthetic overview of Sen's investigation of the interrelationship between rationality and freedom both with respect to individual and collective choice. Of course, these two aspects of decisionmaking are entangled: individuals' rational choices form the basis for collective choice, and norms of rationality are typically thought to be transpersonally binding. Still each may be considered independently as we proceed to assess Sen's volume.

[1] Formerly unpublished material includes the introductory essay "Rationality and freedom" (3–64); "Non-binary choice and preference" (245–58); "Opportunities and freedoms" (583–622); "Processes, liberty and rights" (623–58); and "Freedom and the evaluation of opportunity" (659–712).

Sen has consistently contributed to expanding the boundaries of rational decision theory, famously claiming that were individuals strictly to uphold the narrow axioms of conventional rational choice, they could be described best as "rational fools." This is because a rational agent regarded as either the prototypical *homo economicus*, or the stickler for consistency criteria independent of decision context, is bound to be a "social moron".[2] Sen's claim is historically and philosophically significant insofar as marginalist economists' economic man, who equalized the marginal utility of each last dollar spent on every good, acted to maximize utility in accordance with an objective function analogous to the principle of least action in physics. Sen finds the related ideas of unconscious optimization, and that an individual's life goals can be charted in a single overarching utility function, to be untenable (158–61; 225–8). Furthermore, Sen holds the idea of disembodied consistency criteria as defining a rational choice to be insufficient because in his view the axioms of rationality can only be properly defined in relationship to the specific context of choice (126–32; 225–8). In breaking new ground in his definition of rationality, Sen makes two interrelated claims certain to rankle decision theorists. These two philosophical moves are the acceptance of the standard of "maximization" instead of "optimization," and the acceptance of the context of choice as a central feature of decisionmaking.

"Optimization" requires that an agent select the most preferred, or "best," alternative; "maximization" denotes only that a choice act must not select an outcome that is *known* to be less preferred than another available outcome (746). "Maximization" does not require of rationality one overarching utility function that ranks every alternative against every other; it allows that some alternatives may be unranked *vis-à-vis* each other. It is consistent with the possibility that preference rankings may be dependent upon the "menu," that is the particular range of elements that characterize the set of alternatives available for specific choice acts. Sen's two moves serve to break the dependency of rational choice theory on the premise that agents must have a complete and well-ordered preference ranking across all possible subsets (or menus) of outcomes – that is the same ranking irrespective of the menu or context of choice. Sen further makes the anti-Humean claim that agents are able to reason about ends. Finally, he argues that choice rules themselves are subject to deliberate selection on the part of agents.

For those imbued with standard decision theory, it will not be immediately obvious how Sen defines "rationality," for the reason that his formulation violates the traditional decision-theoretic supposition that the rules of reasoning are somehow unproblematically objective and that

[2] Amartya Sen, Rational fools: a critique of the behavioral foundations of economic theory. In *Choice Welfare and Measurement*. Basil Blackwell 1982, 84–106.

they are therefore orthogonal to the subjective establishment of preference orderings. Sen upholds as fundamental to rationality the principle that "behavior is regular enough to allow it to be seen as maximizing behavior with an identifiable maximand" (30). His understanding of rationality has the following three key features: First, we can reason about what the maximand should be. Second, we can allow the ranking to vary depending on the menu and the context of choice. Third, we can permit incomplete rankings so that given any single menu, what is chosen from that menu is as good as any alternative in the menu, but is not necessarily judged to be at least as good as any alternative in the menu.

In his essays, "Internal Consistency of Choice," and "Maximization and the Act of Choice," Sen articulates his view of rational deliberation as "maximization," in distinction to the "optimization" favored by decision theorists who assume completeness of binary relations among all pairs of preferences, and inter-menu consistency of preference relations (182). Sen tells us that if a decisionmaker had complete preference orderings, with the entailed property of menu-independence, then his specification of maximization as the hallmark of rationality would not differ from the standard choice rule of optimization.[3] "Maximization" contrasts with optimization in a number of crucial ways. Perhaps most prominently, it does not require the weak or strong axioms of revealed preference, which stipulate that rationality inheres in consistency of choice among outcomes regardless of the choice environment. Sen constructs his system to respond to the "need to go *beyond* the *internal* features of a choice function to understand its cogency and consistency" (124). Paul Samuelson's revealed preference approach to consumer choice, which *assumes* context independence and completeness, may conceivably work well for a family of monetarily-based choices. However, it would be insufficient to understand more complex decision situations of productive behavior, collective bargaining, political actions, or consumers' learning (125).

Sen's system can address decisionmaking in two settings that standard expected utility theory is incapable of making sense of: the case in which information is lacking to give the actor a clearly defined optimal choice, and the case in which an appropriate choice hinges specifically on the environment of choice. In the first case, a set of preferences necessarily has an incomplete ordering because, either due to a transitory lack of knowledge or an assertive inability to make a comparative judgment, an agent must make a decision without having a recognizable "best"

[3] To be specific, Sen presents the theorem that maximization is equivalent to optimization if either the preference ranking is complete, or the preference ranking is transitive and there is an optimal choice:

"*B(S, R) = M(S, R) if either of the two following conditions holds: (I) R is complete, or (II) R is transitive and B(S, R) is nonempty*" (183).

alternative. Such is the case, for example, for the donkey that, choosing between two heaps of hay, could become paralyzed either because of indifference between the two stacks, or more significantly because of a lack of knowledge of which is preferable. Sen asserts that in this decision example, maximization is superior to optimization because "maximization will save your life" from paralyzing indecision arising from indifference, and "only an ass will wait for optimization" dependent upon complete information (184, 220–31).

As important as the decision criterion of maximization is to encompassing a choice situation characterized by a lack of knowledge, its more momentous aspect is to be able to encompass the menu-dependent quality that Sen argues is pertinent to many choice acts (168). The context of choice can be characterized by a number of important features that standard decision theory does not adequately address. Having a particular set of choices may in and of itself be significant to a chooser. As well, an outcome selected freely from a fuller set of choices may be less preferred when constrained by external circumstances. For example, a person who chooses to fast for political reasons when given the option of eating well may not necessarily refuse food if no political or moral gesture could be made because the only option is being half-starved. Lastly, information about a particular choice may be conveyed as a function of the particular set of choices among which the agent must choose. A person may decide to have tea with a friend $\{x\}$ given the choice between $\{x\}$ and the alternative of not going $\{y\}$, but spurn the tea-taking alternative if the friend's offer of a third alternative, such as smoking marijuana at home $\{z\}$, makes him reassess the nature of the friend or his home. This structurally not atypical decision scenario stipulates the following pair of menu-dependent choices forbidden in standard decision theory. Given a choice function C(S) (that specifies for any feasible non-empty set S of choices, a non-empty subset C(S) referred to as the choice set of S), in Sen's formal decision theory it is possible that:

$$\{x\} = \mathrm{C}\,(\{x, y\}),$$

$$\{y\} = \mathrm{C}\,(\{x, y, z\})\ (129).$$

Sen argues that these various considerations of menu-dependent choice cannot be resolved by a more encompassing set of preferences that exhibit the prized properties of binary choice and inter-menu consistency. Nor is it the case, as some have argued, that freedom of choice in these context-bounded situations is non-existent for the reason that the chooser has no say over what menu is offered (170–1). For Sen, maximization is superior to optimization because, since it does not require completeness, it can accommodate these irreducible features of decision-making. Rationality, then, acquires an element of free-play because the

agent herself determines what features of a choice environment are relevant, and which in turn structures her hierarchy of preferences and her resulting choice. In addition, as in cases of responsibility for others, the decision-maker has leeway in deciding what decision rules to apply, such as the Savage independence axiom (175–81; 232–9). Thus, it is accurate to conclude that Sen extends the concept of methodological individualism from the subjective acknowledgment of personal preference orderings to the process of identifying the relevant parameters characterizing a choice problem, and establishing which decision rules to apply. In so doing, an "authoritarian" element inherent in assuming *a priori*, as it were, choice axioms dependent on complete and transitive preferences, is removed (6).

Shifting gears to social choice, the complex of issues at the heart of defining individuals' freedom within society has been at the center of Sen's research since his early formulation of the "Paretian Liberal paradox".[4] This paradox pits an individual's right to self-determine ends in direct conflict with the minimalist Paretian criterion for identifying socially superior outcomes. Before proceeding to discuss Sen's more recent contributions to the discussion of collective rationality and rights, it is worth briefly reflecting on the early history of the dialogue between Robert Nozick and Sen.

Human freedom has been a key theme for western society from the time of the Enlightenment to the present, and the Cold War's following hard on the heels of WWII only served to intensify its centrality. As the social choice tradition crystallized in Kenneth Arrow's *Social Choice and Individual Values*, freedom was in some sense built into the theory in the form of methodological individualism (see 300–24). This commitment to the individual as the sole arbiter of social value was entrenched in the assumptions of Arrow's impossibility theorem which accepts that individuals may express any transitive preference ordering, requires that individuals jointly be sovereign over the social choice, prohibits dictatorship, and rules out interpersonal comparisons of the intensity of preferences (329; 591). In this fashion, the methodological assumption that individual choice serves as the foundation of social choice gives primacy to individual preferences uncensored by an external authority; Arrow's conditions for collectively rational decisionmaking rest on nothing but individual choice. We have seen how Sen extends this personal prerogative of choice to encompass contextual factors structuring a decision, and the rules of reasoning themselves (233).

Within social choice theory, methodological individualism, as expressed in the admissibility of any (transitive) individual ordering over social states, is the starting point for deriving a legitimate collective

[4] Amartya Sen "The impossibility of a Paretian liberal". *Journal of Political Economy* 78, 1 (1970), 152–57.

expression of wellbeing. It is to this location of individual freedom within a society comprised of others that Sen turns in order to further explore the interrelationship of rationality and freedom. Within the context of social choice, the individual has no immediate say over collective outcomes; hence, freedom takes on the purely subjective form of expressing one's values but not necessarily securing them in the resulting social outcome. This impasse between subjective preference orderings and collective choice, of course, provided the grist for Sen's Paretian Liberal paradox. Capitalizing on the concept of group decisiveness in social choice theory, Sen shows that a collective decision respecting the Pareto condition of unanimous consent may directly conflict with an individual's right to substantively determine an outcome directly affecting only herself (592–3). The critical insight is that the social choice protocol of respecting individuals' subjective preferences in achieving a collective result directly sanctioned by methodological individualism may still violate the classic liberal value of permitting individuals the freedom to substantively determine outcomes in a personal sphere of action (381–407).

In *Anarchy, State and Utopia*, Nozick's insistence on consequence-independent individual rights was exacting even if moderated by the admission that catastrophic moral outcomes require consideration (638).[5] In return, Sen showed that an all-out commitment to rights may at times lead to catastrophic moral consequences, as in the case of famines wherein individuals cease to be able to purchase food, not for the lack of such a right but due to the lack of an entitlement (86–90; 512). Sen's empirically based rejoinder to Nozick set the stage for accepting the need to evaluate social outcomes in addition to upholding a general emphasis on individuals' rights (639). Upon this platform, Sen concluded that neither a rights-based, nor a social welfarist based approach to social justice, can stand on its own; the two evaluative approaches must be conjoined to build an adequate theory of justice. Sen points out that even the most process-oriented approaches to justice, such as public choice (639–42), Adam Smith's political economy, and Kant's practical reason, still at the end of the day accept the burden of demonstrating that overall social well-being emerges (278–81; 290). Nozick's concession that catastrophic social outcomes must be evaluated is particularly important today because libertarian arguments for free markets tend to rest entirely on process without due attention to social consequences (511). However, it is in adjudicating a delicate balance between seemingly opposed individual rights and social welfare, that the philosophical challenge is at its greatest. It is to this effort that Sen addresses his essays on freedom and collective rationality (e.g., 632–6).

[5] Robert Nozick, *Anarchy, State and Utopia.* Basil Blackwell 1974.

Rights and social outcomes appear to be contrary concerns because rights are defined for individual agents whereas outcomes within social choice theory are defined for groups of agents (due to the interdependence of agents' choices). Furthermore, rights pertain to the process of determining one's ends whereas social outcomes are evaluated in terms of consequences. If the additional concerns of positive and negative liberty are superimposed upon the discussion of rights and social outcomes, the philosophical territory increases in complexity (508–9; 586–7). It then becomes necessary to differentiate between a sphere for private determination of ends, and a sphere in which subjective preferences may be expressed regardless of the resulting social outcome. In other words, within the rights-friendly, negative liberty version of justice, freedom within one's personal sphere of action has substantive hold insofar as within this private arena an individual has the right to make his choice prevail. Within the context of positive liberty, it is more difficult to delimit a sphere of action over which an agent has the purview to determine that an outcome prevail.

Sen advances four major arguments in mediating between a rights-based and social outcome based approach to justice.[6] One is to problematize the ability of the rights-based approach to delimit tidy boundaries around spheres of private action. This is straightforward in the cases of public epidemiology and public information dissemination, both of which enhance individuals' abilities to live and choose freely but require a collective effort (644–6). This difficulty is also obvious in examining the most modest claims of a minimum private sphere for liberty, even in a case so trivial as singing (410–14; 421–6). A second move is to remind readers that much progress has been made in evaluating social outcomes since the time Arrow proposed his original impossibility theorem (65–118). In a related feature to this move, Sen reminds us of his earlier argument that individuals are able to reason about ends, and that therefore public discussion may provide a means to move toward collective agreement about ends as individuals' preferences are potentially altered through the process of debate (286; 290; 587–90). Sen assures us that valid social evaluations can be drawn, even given the incompleteness of individuals' preference orderings (611). Sen argues, third, that game theory is an invaluable tool for understanding the consequences of individuals' exercise of rights, but that ultimately a social judgment over the benefits of this joint exercise is inescapable (311–18; 428–31; 439–60). Fourth, Sen refocuses the discussion from the welfarist notion of individual wellbeing

[6] "The possibility of choice choices" (65–118); "Individual preference as the basis of social choice" (300–24); "Minimal liberty" (408–38); "Rights: formulation and consequences" (439–60); "Markets and freedoms" (501–30); "Opportunities and freedoms" (583–622); "Processes, liberty and rights" (623–58); and "Freedom and the evaluation of opportunity" (659–95).

to the terminology of opportunity. For Sen it is key that "more freedom gives us more *opportunity* to achieve those things that we value, and have reason to value" (585).

Given these four sets of arguments reconciling rights with ends achieved, the final set focusing on opportunities brings us back to Sen's signature contribution to social choice theory, a contribution that is enriched by formerly unpublished essays. In the form of his "capabilities" approach, Sen makes freedom central to any evaluation of social ends in a quantifiable manner of relevance to a public policy analyst, or even a World Bank official (e.g., 658–95). Sen incorporates freedom into social choice theory by assessing individuals' opportunity sets. Preferences over outcomes, such as commodity bundles, miss what is of key importance for Sen: that individuals are deeply concerned with what substantive opportunities are available to them. Thus far, within economic literature the technical language to differentiate between an agent's access to concrete achievements instead of to merely goods has been less developed. To address this deficiency, in granting that the familiar welfarist language of "individual wellbeing" is discredited either for relying on inter-personal comparisons of utility or for proposing objective needs, Sen reconstructs the basis of social choice on human functioning and opportunity sets. That is, the opportunity set an individual is presented with is as important to evaluating his freedom as is his autonomy in decisionmaking and freedom from external interference. Sen argues that individuals with physical disabilities are doubly hampered if they (1) have low incomes because of disabilities, and (2) then can function less ably than a healthy individual with the same income because of the cost of medications or prosthetics (512–27).

In arguing that freedom consists in the ability to realize self-determined ends, Sen incorporates a substantive claim into his analysis of freedom: An agent's freedom is directly linked to what opportunity he has to realize his ends. Of course, this substantive component will alarm negative liberty theorists who uphold the principle of non-injury while eschewing that of mandatory beneficence. Sen extends the market-friendly concept of Pareto efficiency characterizing competitive equilibrium, assessed in terms of individuals' preference fulfillment, to weak Pareto efficiency of competitive equilibrium, evaluated in terms of opportunity-freedom; he establishes that "any competitive market equilibrium is weakly efficient in opportunity-freedom (in a standard commodity space)" (522). He suggests, without direct articulation, that due to the doubly-hampering effect of lack of income upon goal realization if combined with lack of functioning, a slight diminishment in a well-endowed, healthy individual's command over commodity bundles (a) may negligibly affect her opportunity-freedom, and (b) may provide the basis for significantly increasing another individual's opportunity-freedom. In *Rationality and*

Freedom, Sen's goal is not to assign responsibility for this type of beneficent redistribution so much as it is to specify that "the challenge the market systems have to face must relate to problems in equity of distribution of substantive freedoms," and not simply distribution of commodity bundles (526).

Not only do the essays in *Rationality and Freedom* provide a synthesis of the state of the art in social choice theory and axiomatic decision theory, but they are a tremendous aid in understanding Sen's myriad, subtle and profound contributions in these fields. This review has not done justice to the reach of material presented in this text; essays on "Non-binary choice and preference," "Positional objectivity," "On the Darwinian view of progress," on "Environmental evaluation," and "The discipline of cost-benefit analysis" have not been mentioned. I eagerly await *Freedom and Justice*, the companion volume, to which Sen alludes in his essay "Processes, liberty and rights" (623–58) for further insights into his ambitious project of clarifying and exploring the complex interrelationships between human rationality, individual freedom, and social justice.

S. M. Amadae

New School University

DOI: 10.1017/S0266267104230310

Learning and Coordination: Inductive Deliberation, Equilibrium, and Convention, by Peter Vanderschraaf. Routledge 2001, xx + 222 pages.

Peter Vanderschraaf's book is about the problem of how people manage to coordinate their actions. The simplest example is the well-known one of choosing a side of the road to drive on. When two cars approach each other they can safely pass each other either by each choosing the right or by each choosing the left side of the road, while they will collide if one car chooses right and the other left, or vice versa. When they both choose left or right, their actions are coordinated. Car drivers could achieve such coordination case by case, by stopping their cars, taking counsel with one another and agreeing on a side they will each choose. However, such case-by-case coordination would make social life a tedious business and probably a dangerous business as well. In real life we manage to achieve coordination through the establishment of conventions. Car drivers use the right side of the road as a rule. Or the left sides, of course, since it is exactly a characteristic of such situations that there are several possibilities that are on a par as regards people's preferences. Settling on an outcome involves a break of symmetry.

The existence of conventions poses a number of interesting problems for social scientists. How, for instance, do conventions emerge? In which circumstances can we expect conventions to emerge? Which conditions must be satisfied in order for a convention to be possible, and stable? Such questions have proved remarkably difficult to answer. The customary way of dealing with them is through the technical apparatus of game theory, and this is also Vanderschraaf's approach. Within the language of game theory, the situation translates to the problem of equilibrium selection. The situations where social agents can either coordinate or miscoordinate their actions are modelled as games that have several equilibria in pure strategies. The problem the agents face is to find their way to one of these equilibria. Note that in a game where coordination is the main problem the pure-strategy equilibria – corresponding to strategy combinations that are candidates for conventions – are all Pareto-optimal outcomes, but need not be equally desirable for all agents. In the well-known 'Battle of the sexes' game, for instance, with the pay-off matrix shown below, there are two pure-strategy equilibria, (u, l) and (d, r), but R definitely prefers the former to the latter while C prefers the latter to the former. Both however, prefer either of these two outcomes to all other ones, and this is what characterizes the problem situation as a coordination game.

		C	
		l	r
R	u	2, 1	0, 0
	d	0, 0	1, 2

The book is divided in two parts. One, consisting of the first and fourth chapter, discusses the general problem of coordinating the actions of several people and the emergence of conventions to achieve this from a game-theoretical perspective. The other part, consisting of the second and third chapter, is of a more technical kind. They discuss various notions of game-theoretic equilibrium that are relevant to coordination games, and a dynamics that enables agents in such games to reach an equilibrium outcome. In the fourth chapter Vanderschraaf aims to apply the technical results developed in the middle part to the coordination problems introduced in the initial chapter. However, and this is the main weakness of the book, the relevance of many technical results to the general problem is rather limited.

Vanderschraaf's overall aim is to emphasize the significance of correlated strategies for the analysis of coordination games and the emergence of conventions. Strategies of different players are correlated if the probability that one player performs action X is not independent of the probability that another player chooses strategy Y. The above game of 'Battle of the sexes' can serve to illustrate this notion. Apart from the

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

two pure-strategy Nash equilibria (u, l) and (d, r), there is also a so-called Nash equilibrium in mixed strategies, or mixed Nash equilibrium for short, where R plays u with probability $2/3$ and d with probability $1/3$ and C plays l with probability $1/3$ and r with probability $2/3$. Instead of the sure payoffs of the two pure-strategy equilibria, the players now each reckon with an *expected* payoff of $2/3$, which is for both players less than the payoff of either of the two pure-strategy payoffs. This is not surprising because, when playing this equilibrium, almost half of the time the two players miscoordinate, playing (u, r) or (d, l), due to the fact that they pick their strategies independently. If they manage to correlate these choices, however, such that R plays *u if and only if* C plays l, and R plays *d if and only if* C plays r, they will avoid the miscoordinations. It would seem R and C will need some mechanism to establish the correlation between their strategies. They could, for instance, flip a coin before picking their choices, or watch the flipping of a coin by someone else, and play (u, l) if the coin lands heads and (d, r) if the coin lands tails. But according to Vanderschraaf, the playing of a correlated equilibrium can be achieved even without the occurrence of a random event to which the players attune their strategy choices. If each player hypothesizes that the other players correlate their strategy choices among each other, and chooses his or her own strategy to match that hypothesis, then some of these hypotheses can be seen to form an *endogenous correlated equilibrium*. This form of equilibrium makes no sense for two-person games, since there each player faces only one opponent, who cannot be hypothesized to correlate strategy choices with anyone. But take the following three-person game:

		C	
		l	r
R	u	6, 6, 6	2, 7, 2
	d	7, 2, 2	0, 0, 0

f M s

	C	
	l	r
u	2, 2, 7	0, 0, 0
d	0, 0, 0	0, 0, 0

Here R chooses the up or down row, C the left or right column and M the first or second matrix. Each of the three players is now free to believe that the other two players are correlating their strategy choices. Suppose that R assumes that C and M play (l, f) with probability $2/3$, (r, f) and (l, s) each with probability $1/6$ and (r, s) with probability 0, and suppose that C has similar beliefs, *mutatis mutandis*, about the joint strategy choices of R and M, and M about the joint strategy choices of R and C. These beliefs are in equilibrium in the sense of a Nash equilibrium of mixed strategies: for none of the three players their best response given their beliefs is inconsistent with what the other players assume about him or her. What makes such equilibria attractive is that a player's expected payoff – expected on this player's subjective beliefs – can be greater than

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

the expected payoff of the "solutions" that have traditionally been studied. In the above game the expected payoff of R, C and M, given their beliefs, is 4.67, which is greater than their expected payoff of 4.48 for the mixed Nash equilibrium where R, C and M are expected to play u, l and f resp. with probability $4/5$ and d, r and s resp. with probability $1/5$.

The notion of endogenous correlated equilibrium is Vanderschraaf's main technical contribution to the literature on game solutions. It combines the possibility of playing correlated strategy choices, introduced into game theory by Aumann, with an interpretation of mixed strategies as strategy choices that are not actually randomized, on the basis of a particular chance mechanism, but are being conceived as randomized by the other players. It can be questioned, however, whether the concept of endogenous correlated equilibrium makes sense. The choices that players in a game settle upon form an equilibrium if, in case the players know or would come to know the strategy choices of their opponents, none of them would have an incentive to change his or her own choice of strategy. In this way the equilibrium can be seen to fix the strategies of the players. In an endogenous correlated equilibrium, however, no strategies are fixed at all. Take the three-person game introduced above and suppose the players' beliefs are as specified. On these beliefs, each player is indifferent concerning a choice between his or her two available pure strategies. In order for the equilibrium notion to do any work for the players, moreover, they should know each other's strategies. To ensure this, the players' beliefs are assumed to be common knowledge among them, additional to the standard assumption of common knowledge of the structure of the game and the rationality of the players. But on this assumption, each player can derive that all players can derive that none of them has any reason to pattern his or her strategy choice in whatever way. I think it is fair to say that, on these common knowledge assumptions, the players must expect anything to happen. No particular probability distribution of strategy choices is defensible over any other and the notion of expected utility loses all foothold. It seems that players who find themselves at an endogenous correlated equilibrium have gained absolutely nothing.

Endogenous correlation also runs into difficulties with Vanderschraaf's dynamics. During a process of dynamic deliberation, the players update their beliefs about each other during a number of "rounds" and according to a specific updating rule. The rule used by Vanderschraaf is the Dirichlet dynamics: the probability that a player i attaches to another player j playing a strategy s_{jl} increases as a function of the number of times s_{jl} was j's utility-maximizing choice during previous rounds, the total number of rounds played, i's initial probability that j would choose s_{jl}, and a parameter signifying the "boldness" with which i updates his or her probability assessments. The rule can be extended easily to allow for the possibility that players correlate their strategy choices.

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

There are two interpretations of this process of deliberation (although Vanderschraaf actually distinguishes three, but as far as my remarks are concerned, I can lump two of them together). One interpretation sees the updating as done by real players who repeatedly play against each other. At each round they pick a strategy that is utility-maximizing given their expectations about the other players' choices, and after each round they update these expectations on the basis of their observations of the strategies the other players chose. No common-knowledge assumptions are necessary for this process to work. The players do not need to know anything about each other except the actual strategies chosen in each round. If all players are updating their expectations according to a Dirichlet rule, then their strategy choices will, in almost all cases, converge to an equilibrium of one kind or another, or so it is claimed. In the other interpretation, the so-called "fictitious play" interpretation, the updating process supposedly describes the reasoning players go through during a single play, prior to their choice of a strategy, on the basis of an initial set of expectations about each other. A player takes into account both her own maximizing choice and the choices that are maximizing for all other players. This interpretation requires the maximum of common knowledge assumptions, i.e., the structure of the game and the payoffs, the character of all players as both expected-utility maximizers and Dirichlet updaters of a particular kind – meaning that they take the possibility of correlations among other players' strategy choices into account and that they show a particular measure of "boldness" – and finally all players' initial probability assessments of other player's strategy choices. Each player mimics, as it were, all rounds of the game that would be played, were players with identical beliefs about other players actually playing a sequence of repeated games. As soon as the process of deliberation has converged to a stable situation, such that another round of simulated play no longer leads to changes in the strategy choices of the players, a player concludes that her own strategy choice should be her contribution to this simulated equilibrium.

Will Dirichlet deliberators ever arrive at an endogenous correlated equilibrium, even if arriving there does not help much because it fixes no-one's contribution? I doubt it. For a start there are some relatively minor difficulties. The Dirichlet updating process is exact only in rational numbers. For any combination of utilities, prior probabilities and levels of boldness that leads to a mixed or endogenous correlated equilibrium it is the case that a very slight change in one or more of these values will no longer lead to that equilibrium. In fact, for any updating process that does not deal in rational numbers, as will generally be the case for computer simulations and for material deliberators, endogenous correlated equilibria, including ordinary mixed equilibria as a special case, are unreachable. These blemishes can be repaired, for instance by

introducing threshold parameters (which are to be common knowledge) such that utility differences smaller than a player's threshold value are ignored, but I find it deplorable that Vanderschraaf does not discuss this matter at all.

A more serious problem is the way Vanderschraaf makes use of the Dirichlet rule in order to lead players to an endogenous correlated or mixed equilibrium. As soon as players' updated beliefs have reached values that correspond to such an equilibrium, where all of their (undominated) pure strategies become equally maximizing, Vanderschraaf has them change over to a subtly but crucially different rule. The result of applying this modified rule is that the updated probabilities for other player's strategy choices are equal to the values of the previous round. This guarantees that, once players have reached the beliefs corresponding to a mixed or an endogenous correlated equilibrium, they will retain these beliefs forever, irrespective of the strategies their opponents actually pick (in repeated play) or are imagined to pick (in fictitious play). But this scenario can only make sense for fictitious play, since this variant of the Dirichlet rule can be applied only when a player *calculates* another player's *expected utilities* for that player's various pure strategy options. Only in this case does a player know when that other player has reached a point where she becomes indifferent between these options. In repeated play, on the other hand, a player only observes the *strategy chosen* by another player, and assumes this strategy was the utility-maximizing one, given that player's unknown probability assessment. A player cannot *observe* that another player's probability assessment leaves that player indifferent between her pure-strategy options. In the repeated-play interpretation, therefore, players have no use for Vanderschraaf's modified Dirichlet rule. As a consequence, in game sequences where players update on the basis of their observations of actual strategy choices by their opponents, convergence occurs only to a pure-strategy Nash equilibrium, irrespective of the way a player settles on a particular strategy when he or she is indifferent.

Endogenous correlated equilibrium can therefore be reached only by perfectly rational agents who are quasi transparent to each other. But the point of introducing a dynamics in the first place, in my view, is to model the process by which actual social agents can arrive at coordinated action and the establishment of a convention. Fictitious play between transparent agents seems a definite non-starter for our understanding of this process.

What undercuts the Dirichlet dynamics still further is that the problems mentioned carry over to the establishment of exogenous correlated strategies, i.e., strategies correlated on the outcomes of particular public happenings, such as coin flips, dice throws, sunspot appearances or what have you. In fact, if all players in a sequence of repeated games are modelled as Dirichlet deliberators, there are no exogenous correlated

strategies, only endogenous ones. Because none of the players is paying attention to the external event in order to determine his or her own strategy choice, but only to see if the way it comes out is linked to a pattern in the choices of the other players, what results is a collection of endogenous correlated equilibria, one for each way the dice, or whatever it is, may fall. And as discussed, unless we allow players to be transparent, in the long run such an equilibrium can only be a pure-strategy Nash equilibrium. If, on the other hand, some of the players are modelled not as Dirichlet deliberators but as players who in fact base their strategy choice on the result of the public happening instead of on their expectations of the remaining players' strategy choices, then the remaining Dirichlet players will equally drive the outcome, for each of the external states, to a pure-state Nash equilibrium (if any exists).

It is a bit surprising, with the amount of attention endogenous correlated equilibrium receives in the technical chapters, that it is completely absent from Vanderschraaf's discussion of the nature of conventions in the final chapter. What he emphasizes there is that conventions can take the form of correlated equilibria, while the standard literature discusses only pure-strategy equilibria. In order to make room for correlated equilibria, he modifies the definition of convention given by Davis Lewis, whose 1969 study was the first to approach the problem of coordination and conventions from a game-theoretic perspective. Since his view that conventions can take the form of correlated strategies is in itself well taken, it is a pity that Vanderschraaf does not discuss some examples of such conventions. All the more so since a simple coordination problem he mentions in the introductory part of the book, namely "Telephone Tag" – who calls whom back when a telephone conversation is accidentally interrupted? – seems actually to have been solved by a rule that has the form of a correlated equilibrium. Another example is the giving of way at crossroads: both the rule that one gives way to traffic approaching from a particular direction (the right in most countries) and the rule that red traffic lights mean "stop" and green ones mean "go" can be seen as conventions that are correlated equilibria.

So one must finally conclude that the book's technical results hardly support its general message. The topic, however, remains an important one, and the intricacies of correlating the strategy choices of social agents in game-like situations are an interesting line to follow. It is to be hoped, therefore, that Vanderschraaf will find the occasion to write the sequel to this book that he mentions in the preface.

Maarten Franssen

Delft University of Technology

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

DOI: 10.1017/S0266267104240317

Hansson, Sven Ove, *The Structure of Values and Norms*, Cambridge University Press, 2001.

Hansson has written an excellent book on the logic of preferences and norms. In it, he both illuminates the concept of preference through logical analysis, and connects it to value predicates like "good" or "worst" as well as to normative predicates like "should" and "ought." Although formal in style, the book is by no means written for logicians only. Hansson takes great care to discuss the intuitions behind the formal framework and strives for a compromise between realism and formal rigor. Anyone interested in economics, decision theory, political philosophy or social choice theory is well advised to familiarize herself with the (not too difficult) logical machinery, as there are lots of insights to be reaped from Hansson's work. The book incorporates reworked material from 18 of the author's papers, written over the last decade and a half. In addition, it provides many new results due to its unified approach and to Hansson's often critical scrutiny of his earlier views.

The book consists of three parts: a discussion of preference relations, and, building on that, a discussion of monadic value predicates and of norms. The first part is the most extensive and is fundamental for the other two. It is also the one where Hansson's work is most innovative. Here he introduces a new justification for the rationality requirements imposed on certain types of preference and compares it with the mainstream money-pump arguments. He then offers a model of preference change completely different from previous approaches and finally proposes a *holistic* interpretation of non-basic preferences in terms of basic ones. With the very sensible distinctions that Hansson makes between different types of preferences, this section offers an interesting analysis of the necessary conditions for understanding preferences, with important results for decision theory. In the second part of the book, Hansson seeks to define monadic value predicates like "good" or "worst" in terms of the relational predicates "better" or "worse." In the third part, he discusses normative predicates like "should" and "ought." He distinguishes these predicates according to their range of applicability: whether they refer to a particular and actual situation, to a counterfactual or to a general context. His innovative contribution here consists mainly in the development of an alternative semantic for these predicates, and in his attempt to define them in terms of preferences, while carefully avoiding a reductive account. In the final sections of the third part, he uses general normative predicates to analyze legal relations.

In the following, I will restrict my discussion to four aspects of Hansson's rich book: the rationality requirements imposed on different

types of preference, his model of preference change, his account of holistic interpretation and his treatment of situationist and counterfactual norms.

Rationality Requirements. Hansson makes two basic distinctions between preferences. First, he differentiates preferences with respect to the types of their alternatives. An agent might prefer one alternative to another, she might be indifferent or withhold her judgment. The alternatives the agent compares in this way are descriptions of events or facts; they therefore are susceptible to logical inference. If the alternatives are logically inconsistent, Hansson calls the preference relation *exclusionary*, if the alternatives are logically compatible, the preference relation is called *combinative.*

The second distinction Hansson makes between preferences is with respect to their function. A preference minimally functions as a comparison between two alternatives. Such a preference is called *pairwise.* But the agent might see her comparisons in a larger context of many alternatives, as one does when having a preference between wine, beer and juice for a drink accompanying dinner. Here, the combination of pairwise comparisons helps in making a choice from a set of alternatives larger than two; preferences which fulfil such a function are therefore called *choice-guiding.*

The only requirement that pairwise exclusionary preferences have to satisfy is the reflexivity of weak preferences. Choice-guiding exclusionary preferences, on the other hand, have to satisfy more restrictive rationality criteria. Nonetheless, it is remarkable how weak Hansson's minimal requirements for rational preferences are: neither completeness nor full transitivity are stipulated. Beyond these minimal requirements, Hansson sees the question of rational preference in terms of a trade-off between function and cost. "Is it more costly to make my preferences more sophisticated or to deliberate with a rudimentary preference ordering?" is the consideration that determines the choice of criteria. In particular, it determines the extent of the preference ranking over the set of alternatives, as well as over subsets of this set. Having to choose between three brands of tomato ketchup, for example, the agent might prefer brand A to both B and C. A preference relation between the latter two does not help her decision at that moment, hence establishing this preference does not yield a net gain for her. Nevertheless, if she anticipated that brand A might be unavailable in the future, then establishing a preference relation over the alternative set $\{B, C\}$ might yield an advantage for her future choices that is greater than the comparison costs involved. Depending on this trade-off, it can be rational to require the preferences to satisfy transitivity, completeness or antisymmetry.

Hansson's decision-theoretic approach is an innovative alternative to the common justification of the rationality requirements imposed on preference sets, but it is not without controversy. For want of an objective

measure, the "gain" or "loss" from making one's preferences more or less sophisticated is comprehensible only as the agent's subjective evaluation. She decides how much comparison costs she is willing to trade for more versatility in future decision situations; and she does so on some sort of meta-ranking. Thus Hansson's approach is in danger of infinite regress, since the rationality criteria for an evaluation would then be justified by recourse to another evaluation. The justification must therefore be kept on the informal and intuitive level, and our intuitions as to what is advantageous and what is not, might just be as vague as our direct intuitions about rationality requirements were in the first place.

Preference Change. If a choice-guiding preference ranking must satisfy certain rationality criteria, then the change of a single pairwise relation in it, or a change in the alternatives available, might have consequences for the whole ranking: to maintain consistency with the criteria, the ranking has to be adapted to the change. This is the basis of Hansson's model of preference change, which is closely related, both in structure and in its proof methods, to the better known model of belief revision of Alchourrón, Gärdenfors and Makinson. The three pillars of his model are: (i) the interpretation of preference changes as initiated by one of four *inputs:* the change and the removal of a binary relation, and the addition and the removal of an alternative; (ii) the reducibility of all inputs to sequences of these four basic types; and (iii) a sentential representation. The first assumption, in particular, is problematic:

> the input-assimilating model is based on the simplifying assumption that the cause(s) of a change can be represented in the form of an input. (44)

In belief change, the intake of new information is the predominant if not exclusive cause of new beliefs. At the same time, the sentence representing this information can be used as the input that (logically) necessitates further adjustment in the belief set. The causes of belief change and the representation of beliefs are thus closely connected. In the case of preference change, the cause and the input often come apart. The causes of preference change are manifold: social pressure, new beliefs, physical conditions, etc. The input in Hansson's model, on the other hand, is only a command to change or remove a relation or an alternative in the representation. The connection between causes and Hansson's input commands would require a separate causal model, but such a model does not fall within the competence of a logical treatment. What is not clear is whether all causes of preference change can even be represented as the input that Hansson requires. For example, social pressure might have an effect through a change of the consistency criteria imposed on the preference model (e.g. by telling the agent to "loosen up" or "enjoy a

little craziness" in her desires). For such a case, no "input" in Hansson's sense can be constructed.

Unlike the common belief dynamics models, Hansson's model of preference change starts with a *revision operator.* Two principles guide its construction: the preference adjustment should be minimal, and which preferences are adjusted depends on some external information – the so-called priority index. Hansson's way of designing this priority index is innovative, as he offers a different structure from that used in belief dynamics. The *contraction operator* on the other hand is constructed on the basis of the revision operator. This leads to the satisfaction of the *postulate of recovery*, which states that a preference model contracted by a preference can always be recovered by revising the contracted model by exactly that preference. The controversial character of the recovery postulate is revealed in the following example: an agent prefers A over B and B over C. Hence by transitivity, she prefers A over C. Now she drops her preference $A > C$. In order to comply with transitivity, at least one of the other two preferences has to be removed from her overall evaluation (and it might well be possible, for lack of a specifying criterion, that she removes both). In any of the three resulting versions, a subsequent revision by $A > C$ will not restore the original preference model.

Original preference model: $\{A > B, B > C, A > C\}$			
Contraction by $A > C$:	(i) $\{A > B\}$	(ii) $\{B > C\}$	(iii) $\emptyset$
Revision by $A > C$:	(i) $\{A > B, A > C\}$	(ii) $\{A > C, B > C\}$	(iii) $\{A > C\}$

Models of preference change should allow for such cases, as they play an important role in preference dynamics. The recovery postulate is therefore overly restrictive.

Preference Holism. Hansson next turns to combinative preferences, which are differentiated from exclusionary preferences by the structure of their alternatives in two ways. First, combinative preferences have logically compatible relata (like "I prefer owning a flat in New York to owning a house in Tuscany"), while the alternatives of exclusionary preferences are mutually exclusive ("I prefer being a student over not being a student"). Second, alternatives of exclusionary preferences are maximally specified, while relata of compatible preferences are not so. Hansson distinguishes two ways of relating exclusionary to combinative preferences. The *aggregative approach* derives exclusionary preferences from combinative preferences. The *holistic approach*, on the other hand, takes maximally specified alternatives as the fundamental bearers of value and interprets combinative preferences with reference to them. Hansson, who subscribes to the second approach, constructs an interpretation of pairwise combinative preferences in two steps. In the first step, he

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

offers an interpretation of combinative preferences as a relation between incompatible alternatives. The basic idea is to reinterpret a combinative preference for p over q as a preference for $p \wedge \neg q$ over $q \wedge \neg p$. A problem arises in cases where it is logically or causally impossible that $p \wedge \neg q$. Hansson casts such a possibility as the case where for all maximally specific alternatives $A \in \mathcal{A}$ such that $p \in A$, it necessarily holds that $q \in A$ and notates this $p \models q$. His amended translation procedure for combinative preferences then runs as follows. First he defines:

> $p/_{\mathcal{A}}q$ ("p and if $\mathcal{A}$-possible not-q") is equal to $p \wedge \neg q$ if $p \models q$ is false. Otherwise, $p/_{\mathcal{A}}q$ is equal to p.

And further:

> The informal statement "p is better than q" is translated into $(p/_{\mathcal{A}}q) > (q/_{\mathcal{A}}p)$, and "$p$ is equal in value to q" is translated into $(p/_{\mathcal{A}}q) \equiv (q/_{\mathcal{A}}p)$. (70)

With the help of the translation procedure, Hansson in the second step constructs a selection function f from pairs of (interpreted) combinative alternatives to pairs of maximally specific alternatives. This way, combinative preferences can be interpreted with reference to preferences over the set $\mathcal{A}$ of maximally specific alternatives:

$$p \geq_f q \text{ if and only if } A \geq B \text{ for all } \langle A, B\rangle \in f(\langle p/_{\mathcal{A}}q, q/_{\mathcal{A}}p\rangle).$$

What is at issue here is which pairs $\langle A, B\rangle$ the function picks out. Hansson proposes the *ceteris paribus* approach:

> Any given alternative that contains $p/_{\mathcal{A}}q$ is preferred to any alternative that differs from the first in that it contains $q/_{\mathcal{A}}p$, but is otherwise as similar as possible to it. (75)

Hansson then correctly points out that a *logical* operationalization of the concept of "as similar as" can only work under assumption of logical atomism and hence should not be followed. Instead, he employs a four-place similarity relation (on the basis of extralogical information) and discusses its logical properties, but abstains from providing any clue as to how it could be measured.

My concern here is with the *ceteris paribus* approach. If the maximally specific alternatives are broad enough, for example if they were possible worlds, hardly any combinative preference would stand the *ceteris paribus* approach. Take an example of Rainer Trapp: even though it can plausibly be said that I prefer contracting Cholera to having cancer, I prefer the second to the first in a world where Cholera was incurable but cancer was curable. Hansson himself quotes this example, but claims that his approach of *Myopic Holism* avoids this problem. Myopic Holism takes

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

maximally specified alternatives to be "alternatives that cover all the aspects under consideration – but not all the aspects that might have been considered" (59). This introduces an arbitrariness which puts the formal rigor into question – how can we be sure, as Hansson seems to be, that "in general, the maximally similar but contextually irrelevant pairs of complete alternatives have been excluded when the alternative set was selected" (78), and why should we rest content with such a vague selection process? Hansson here jumps on the *small-world* bandwagon without clarifying the question of what these small worlds are supposed to be.

The *ceteris paribus* approach was designed to interpret pairwise combinative preferences. When interpreting choice-guiding combinative preferences, in contrast, all maximally specific alternatives which represent a combinative relata have to be taken into account. Hansson therefore interprets choice-guiding preferences as preference relations over sets of maximally specified alternatives. For example, "I prefer staying home over watching a movie" is a preference for the set of possible domestic experiences over the cinematic ones. Sometimes, information is available as to which of the alternatives in one set will be realized. Depending on that information, Hansson distinguishes two approaches. The *prognostic approach* uses all available information, while the *agnostic approach* treats the outcome as completely undetermined. Let's imagine a choice between staying home and going out to watch a movie, when the two things to do at home are either practicing the piano or watching TV and the values of the three alternatives are $V(piano) = 4$, $V(cinema) = 2$ and $V(TV) = 0$. The prognostic approach determines the preference between staying home or going out as a weighted average of the values of the maximally specific alternatives. If it is probable that if I stay home, I will watch TV, then I will prefer going out.

It is noteworthy that this approach does not really fit with the rest of the book: in order to obtain a weighted average, cardinal values are needed for the alternatives, and this cardinal information cannot be derived from the exclusionary preference relation alone.

Norms. In the last part of his book, Hansson attempts to show that deontic predicates can be defined on the basis of preference relations. First, he claims that all action-focused normative predicates can be translated into a predication of states of affairs. "The agent ought/is permitted/must not **a**" becomes "It is required/permitted/prohibited that the agent do *A*." This way, normative predicates take the same arguments as preferences. Second, he requires normative predicates to differ in stringency, such that they are ranked according to their strength. Third, he requires the three groups of predicates to be *unilaterally* interdefinable at all levels of stringency. That is, for "It is wrong to do X" at any level, there

is an equivalent predicate "It is required not to do X;" and similarly for "It is permitted to X," which is equivalent to some predicate "It is not required not to do X." Fourth, he distinguishes between different "perspectives" of one and the same situation, with a perspective being "that which determines what states of affairs are taken into consideration in the appraisal of a given situation" (135). Fifth, Hansson distinguishes normative predicates by the situation they refer to: to a particular and actual situation, or to particular and possible situations. Even though these distinctions, in particular the fourth and the fifth, sound plausible at first hearing, Hansson makes too heavy use of them to rely on their intuitive justification alone. A more formal treatment would have been desirable here.

Situationist deontic logic treats normative predicates that refer to a particular and actual situation within one perspective. Hansson finds fault with the basic semantic structure of standard deontic logic for its principle of *necessitation:* "whatever is necessitated by a moral requirement is itself a moral requirement" (141). Hansson holds this principle responsible for the so-called deontic paradoxes (Ross, Good Samaritan, etc.). He instead offers an alternative semantics for deontic logic, by stipulating prescriptive predicates to be *counternegative.* Counternegativity relates the validity of a normative predicate to a combinative preference ordering in the following way:

> H is counternegative with respect to a given preference relation $\geq$ iff $Hp \wedge (\neg p) \geq (\neg q) \rightarrow Hq$

As the predicates are interdefined, this provides the core for a semantics for all groups of predicates. Hansson then proceeds to show that the nondesirable properties of standard deontic logic do not hold for most of those predicates (that is, mainly versions of the necessitation principle), while desirable properties like *agglomeration* and *permissive cancellation* do hold.

I do not find Hansson's alternative altogether convincing, because it substitutes a system with too strong principles for something too weak. After all, necessitation is plausible in many circumstances, and should not be altogether dispensed with. Take the example where you ought to support your grieving friend, and supporting him implies that you do not go on a planned holiday that day. Then it seems correct to conclude that you ought not to go on that holiday – but nothing in Hansson's situationist logic allows such a derivation.

Next, Hansson extends the normative predicates to counterfactual situations. These applications are primarily of interest for their possible violation of the "ought implies can"-principle. The case where prescriptive predicates cannot be obeyed leads to a *moral dilemma.* Hansson's treatment is twofold: on the one hand, he preserves the normative predicate as

https://doi.org/10.1017/S0266267104240317 Published online by Cambridge University Press

a moral but non-obeyable requirement, on the other he "pragmatically resolves" (175) the situation by introducing an action-guiding, *maximally obeyable* prescription derived from the moral obligation. This is an important distinction: you are not free to do whatever simply because you have violated an obligation – a problem that arises in standard deontic logic. Unfortunately, Hansson offers a formal treatment only for that subset of counterfactual predicates whose antecedent removes some of the alternatives of action.

In conclusion, despite my criticism, I think this is an excellent book. Preference logic is a relatively new field, and Hansson certainly sets its newest standard. Any research on those topics will have to consult his work, and researchers will find the many properties proved for the concepts, as well as the proofs themselves, of great help.

Till Gruene

London School of Economics

DOI: 10.1017/S0266267104250313

The Theory of the Individual in Economics: Identity and Value, by John B. Davis. Routledge, 2003, viii + 216 pages.

Any theory that apparently shows how the exercise of individual freedom can lead to efficient outcomes rather than some form of chaos is bound to appeal to a political theory that places individuals and their freedom at its center. This is not the only source of attraction between neoclassical economics and liberal political theory. Political arguments are naturally stronger when they seem to work with a recognizable model of human behavior, and the apparent power of the rational choice model in explaining politics has as a result further deepened the relationship.

The point of John Davis's new book, however, is to signal the end of this affair. The reason is simple. The picture of the individual supplied by neoclassical economics is too slight to justify the weight placed on the individual in liberal political theory. A different, "thicker" model of the individual is required for this purpose: one where the individual has a recognizable identity which comes from being historically and socially embedded.

The question of what kind of economics should inform political discussion is the broad context for this book. The flip side to this question focuses more narrowly on what kind of individual should be at the heart of economics. Although Davis has one eye on the broader picture, for the most part this is a book for economists and so it is concerned with opening a space within economics where the individual can be discussed

and made weightier, even if this also makes the idea of the individual more problematic. In other words, Davis would like economics to join in the debate that almost every other discipline in the social science and humanities enjoys around the concept of the individual. But, it will only begin to do so once it breaks the habit of treating the individual as exogenous.

Davis draws on heterodox economics for a more substantial conception of the individual. This may seem surprising. This tradition has often been associated with forms of holism and collectivism where the individual seems to disappear. But in an interesting subsidiary argument, Davis explains this apparent paradox through an examination of the historical trajectory of the orthodox and heterodox parts of the discipline. While the mainstream has been slowly destroying the individual in the quest for a particular kind of scientific status, the heterodox margins have been forced to confront precisely the criticism that their emphasis on the social implausibly turns individuals into social dupes. This has driven the heterodox tradition to a more nuanced and substantive account of what the individual consists of and it potentially supplies the 'thicker' notion of the individual which liberal political theory needs. Of course, this will not be the strand of liberal political theory that, say, Milton Friedman, has advanced but it will be a version that has a recognizable type of person at its centre.

The key to Davis's argument is the suggestion that placing the individual at the center of a political theory requires that he or she have a distinct identity. It makes no sense to value the individual qua individual unless there is something about the individual that makes them distinct. Minimally, this means that any theory of the individual that is fit for this role must satisfy two conditions.

First, the "individual" must be distinct from other individuals and other types of agent. This is what Davis refers to as the "individuation" problem. Second the "individual" must be recognizably the same over time. This is the "reidentification" problem and it arises because individuals plainly change over time in some respects (e.g. their endowments, beliefs and possibly even their preferences). So any credible theory of the individual must supply an account of the individual which explains how they are distinct from others and how they have some temporal continuity.

The first part of the book is concerned with showing how neoclassical or mainstream economics fails these two tests. I have so far been a bit loose in referring to neoclassical or mainstream economics when actually Davis draws a clear distinction. Neoclassical economics for him is the version of rational choice theory where there is a clear psychology. People have preferences which motivate them to act and they are instrumentally

rational because they act so as to satisfy best those preferences. Or in folk psychological terms, it is desires and beliefs which enable them to act and identity is subjectively experienced as a kind of inward consciousness. The mainstream differs by dropping the psychology and identifies the individual solely through their choices. This is the behaviorist, axiomatic version of rational choice theory.

Behaviorism fails for two reasons on Davis's account. First, if individuals are simply identified with choices that satisfy the standard axioms then there is no obvious way of distinguishing individuals from other kinds of multi-individual agents, like families, households and firms, because they too are identified with well behaved preference orderings. Of course, if such multi-individual units could be reduced to their individual components, then this would not be a problem. But Davis argues, I think correctly, that such reductionism has not proved possible, at least so far, in economics.

Second, Davis supplies a useful survey of where this kind of behaviorism fits within debates in the philosophy of mind and uses Searle's now famous Chinese room argument as a counter. The purpose of Searle's argument is to show that the correct manipulation of Chinese characters is not identical with understanding or using, in any meaningful sense, the Chinese language. Thus a purely behaviorist account is always going to miss something about individuals: their intentionality. I suspect that a bit more needs to be said for this to be really telling since the point about behaviorism is precisely that it seeks to deal in the observable and intentions are not directly observable. So why should this omission worry behaviorists?

It becomes significant, of course, when the objects of choice cannot be identified independently of the meanings that people attach to them. Or to put this around the other way, it is a problem when the objects of choice are not defined solely in terms of their physical properties. To use another rather famous example, this time from Sen, the apple in the choice set of an apple and a banana may not be the same apple when the choice set is expanded to include another smaller apple. In all physical respects it is the same apple, but it is not implausible to suppose that sometimes they will not be treated as the same by individuals. For example, a person might choose the apple over the banana in the first case but shrink from choosing it in the second for fear of appearing to be greedy by selecting what is now clearly the "bigger" apple. This is an obvious problem for a behaviorist who refuses to use mental state categories because the individual's choices will appear inconsistent when, in fact, they are not. It is just that in the social world, the objects of choice are not simply distinguished by their physical properties. They also sometimes have symbolic properties and this perforce means that the best intentioned behaviorist (are such things possible?) cannot avoid dealing in mental categories.

The argument against the subjective, psychological version of rational choice theory again comes in two parts. Davis rehearses some of the possible ways that an individual might be "reidentified" over time and finds that the most plausible way comes from associating the individual with their preferences. But this creates an evident difficulty when preferences seem to change over time unless one buys some version of Stigler and Becker's "De gustibus non est disputandum" argument. Roughly speaking, this explains apparent preference changes by appealing to a constant utility function while allowing that experience can affect the productivity of a particular action in terms of the utility that it creates. Thus when we observe someone spending more and more time listening to music, it may be natural to assume that their preferences have changed, but there is another interpretation. They could have the same preference for music appreciation, it is just that the experience of music builds a kind of human capital which makes each hour of music generate more music appreciation. Hence the person does not listen to more music over time because their preferences have changed; rather it is because each hour of music now produces more music appreciation than before.

In part, Davis finds this interpretation forced as it seems more natural and simpler to say that a person's preferences have changed. In addition, if Becker and Stigler's interpretation is to be held, then it seems that the individual should be identified not just by their preferences but also by their human capital endowments which affect the quality of experiences like listening to music. But once this is done, since some of these endowments emerge through social interaction of one kind or another, the individual can no longer be taken as exogenous to the analysis.

I am not entirely persuaded of these objections; and nor is Davis to judge from the fact that he takes two important lessons from Stigler and Becker. One is that the individual is, indeed, formed socially in some sense. The other is that the individual should be identified with the kind of capacity to act in the knowledge that action produces changes in oneself; and this is also found in Stigler and Becker.

The psychological version of rational choice also fails the "individuation" test on grounds that come out of the multiple self literature. If we accept that the perception of multiple selves is real enough, then the problem of a unitary self amounts to an intrapersonal version of the familiar interpersonal problem of constructing a social choice rule from individual orderings.

In the second part of the book, Davis moves in a parallel way through the heterodox tradition. First there is a quick tour designed to make plausible the claim that this tradition is the repository in economics for an alternative model where the individual is socially and historically embedded. Most of the usual suspects get a mention here, myself included,

and although we may use different terms, there are typically two ideas that help form the alternative model.

First there is a version of the structure-agency framework which places the individual from the outset in a set of constitutive social relationships. One cannot understand the individual outside of these social relationships on this account as they help define the roles and positions of the individual and these, in turn, give meaning to his or her life. To make the connection here with what is missing from the mainstream model, the meanings of action that are so crucial to the individual's internal world, and which get lost in the model of the mind as a computer, depend for Wittgensteinian reasons on a social existence. There are no private languages which would sensibly give a strictly individually personal voice to the symbolic world that motivates us.

Second there is the idea that individuals are self reflexive. This self reflexivity gives the individual some scope for stepping back from their social maelstrom to make choices. There are choices, albeit limited in some cases, to be made about which social groups to belong to and one can always put a degree of distance between following the prescriptions of a norm and internalizing the values encoded in that norm. These are the spaces where being an individual really begins to mean something that would be worth defending or promoting politically.

Davis then takes a particular model which encodes these two features: one based largely on Tuomela's ideas of collective intentionality. This is a framework where individuals distinctively act on "we intentions" as well as "I" ones. There is a quick aside on why "we intentions" cannot be subsumed within a form of instrumental rationality, but the main point of developing this heterodox model is to put it through the same tests of "identification" and "reidentification" as the atomistic one from the first part of the book.

This model passes the "identification" test because the individual for this purpose is associated with the capacity to self impose "we intentions" and as a matter of definition only individuals can self impose in this manner.

"Reidentification" looks problematic for the embedded model because group affiliation changes over time as does one's position within a group; so use of "we language" alters and this might seem to imply that the individual changes. Indeed, as life seems to become more and more complex the danger of the oversocialized individual pushing out the "self imposer" may seem all too real. Davis admits this, as a matter of practice, but so long as the individual retains a "capacity" over time to self impose then the individual is reidentified through this "capacity." "Capacity" for this purpose is obviously a crucial term and it is used in a manner that is analogous to Sen's use of capability when this describes the opportunities

available to an individual. So the individual has to have the capacity to self impose which comes from having options.

Thus the book engages in an ontological discussion and advances a particular model of a socially embedded individual. It offers an original set of arguments based on the "identification" criteria for the superiority of a socially embedded version of the individual over the neoclassical/mainstream one where, as Pareto famously remarked, the "individual can disappear." It is also well written and covers an immense philosophical domain with observations on the role of ontology, interesting comments on the philosophy of mind, the fact/value distinction and much more. Since I am sympathetic to the need for ontological discussion and to the claims of more socially embedded models of individual rationality, it will come as no surprise that I could hardly put this book down or that I have been recommending it to anyone who comes within air shot.

This said, I want to explore some possible weaknesses in the argument, at least as it stands. None seems to me to be fatal. Indeed, in some degree, they are inevitable in a book that ranges so widely and importantly in such a relatively short space.

First, there is the question of why we should engage in ontological discussion. Ontology has been eclipsed by epistemology in the philosophical discussion around economics and Davis wants to redress this balance. But why?

Davis gestures, not entirely persuasively, to Cartwright on "the importance of capacities as opposed to laws in science" (185). Yet there is a stronger reason for engaging in ontological discussion. Or to put this in a slightly different way, which connects with the famous realism of assumptions debate in the methodology of economics, there is a very good reason why the assumptions in theory about individual agency must correspond in some sense to a centrally recognizable part of individuals in the world. It is because economic theory is used not only to understand the world but also to evaluate and possibly change it.

The point is this. The "as if" defence of assuming that people are preference satisfiers, for instance, may just about work if explanation is the only goal. But the moment the model is used to say that some set of outcomes are or are not Pareto efficient, then the "as if" defence does not work. Such statements depend on individuals *actually* being preference satisfiers.

Related to this, I want to take issue with the neoclassical/mainstream distinction. Davis is not alone in casting the history of orthodoxy in this way (e.g. see Mirowski's latest book), and I quite see that something of this behaviorist sort has been going on. But plainly it is not the whole story. Consider, for instance, the use of efficiency concepts like Pareto efficiency, these terms would be meaningless on a *pure* behaviorist account. The term

involves saying that people are "better" or "worse" off and this needs more than the assumption that their choices satisfy the standard axioms: it requires a psychology. Since I would wager that references to efficiency and Pareto efficiency in the economics literature have not diminished over the last 40 years, I don't think this period can be sensibly characterized in terms of the mainstream turning to behaviorism.

Third, I am not persuaded by the particular claims, as things stand in chapter 7, of Davis's version of collective intentionality. This is not serious in the sense that I could be persuaded, but his model does need to be formalized. There are a range of alternative models of socially located individuals now on offer and formalism helps bring out the significant differences (e.g. Rabin's idea regarding how reciprocal kindness/unkindness can motivate, Sugden's model of normative expectations which do the same and Bacharach's version of team reasoning). Indeed this is actually a rather vibrant area in economics at the moment and since there are genuine issues about how best to capture the normative influence on decision making, we need to bring them out in ways that can be tested.

Finally I would add something to the analysis of identity: our story-telling capacity. What also seems to give us a sense of identity over time is the way that we can tell stories about how we have arrived at where we are and the fact that these stories are personal provide the material for distinguishing one person from another. This is in a sense an observation about how many people experience individuality in the twenty-first century and it does not detract from what Davis says about our capacity to self impose "we intentions" because without such a capacity our stories would be rather sad. But the "narrative self" is one of those useful ideas that economics might get from following Davis's advice and joining the wider debate about individuality in the humanities.

Shaun P. Hargreaves Heap

University of East Anglia

DOI: 10.1017/S026626710426031X

Collective Rationality and Collective Reasoning, Christopher McMahon. Cambridge University Press 2001, vii + 251 pages.

Christopher McMahon's *Collective Rationality and Collective Reasoning* is an examination of two interrelated issues: the way in which cooperation is guided by reason and the ways in which collective reasoning is itself a

form of cooperation. Rational cooperation, according to McMahon, has two aspects: one concerns whether an individual has sufficient reason to contribute to a cooperative scheme (a plan that identifies individual contributions to a collective action) and the other concerns whether a group of individuals who are disposed toward cooperation can regard the choice of a cooperative scheme as rational. The first half of the book provides an account of these two aspects of rational cooperation and considers rational cooperation in the political domain, specifically the democratic state. The second half of the book focuses on the issue of collective reasoning. McMahon argues that the product of collective reasoning should not be understood as a collective judgment concerning what the relevant reasons support. Rather, the product of collective reasoning is a common pool of reasons from which participants can draw to reach their own individual judgments. He goes on to detail the ways in which disagreement in a group can be resolved by collective reasoning. Finally, McMahon attempts to explain some peculiar features of collective reasoning. For instance, collective reasoning, according to McMahon, does not seem to be susceptible to the free-rider problems plaguing other forms of co-operation.

One of the merits of the book is its clear and articulate discussion of a familiar problem in rational choice theory. On standard rational choice theories the principle of individual rationality (PIR) directs agents to perform that action that will produce the most desirable outcome given their preferences (or, as McMahon calls them, "values"). But in prisoner dilemma-type cases this principle directs agents to defect with the result that individuals often get less of what they value than if they behaved "irrationally."

There have been various attempts to solve this problem. Some have tried to solve the problem by pointing to the repetitive nature of cooperative enterprises. Non-cooperation is punished in subsequent repetitions. Agents will realize that what is gained by non-cooperation at a given point is outweighed by what is lost through denial of future cooperative benefits. McMahon argues, and I think correctly, that this won't work. In cases where there are many contributors defectors often go unnoticed and unpunished. Fear of punishment does not eliminate free riders. Others have appealed to the principle of fairness in order to explain why it is rational for agents to contribute to cooperative endeavors. McMahon rejects this as well. To appeal to the principle of fairness might help us in cases where reasons are non-moral, but when an agent has competing moral reasons things get more difficult. The problem seems to be that if the agent has competing moral values, throwing the principle of fairness in just adds another value to the mix. If there are moral values that are in competition with the value of fairness, there is no guarantee that fairness will win out.

McMahon's solution is to avoid appealing to another principle of value and reconsider instead the way that actions and outcomes are correlated. All principles of rationality, according to McMahon, correlate actions with their outcomes, but there are various ways of doing this. According to the PIR an agent correlates each action with an outcome and then rationally chooses the action that is associated with the highest value to the agent. McMahon suggests that the reason to cooperate should be understood in terms of the principle of collective rationality. The principle of collective rationality (PCR) directs individuals to compare the outcomes produced not by their individual action but by two different combinations of actions of which their action is a part. When faced with a prisoner's dilemma an agent following the PCR will not simply consider the outcome produced by his individual act of defection or cooperation but will consider the outcome of the following combination of actions: (I defect, you defect) and (I cooperate, you cooperate). The former combination will produce less value to me than the latter and so the PCR will direct me to cooperate.

Prima facie, this seems to be a nice way of solving the problems facing standard rational choice theories. But it is not at all clear that the solution is a novel one. Susan Hurley (1989), for instance, has argued against a form of individualism that sounds very similar to McMahon's account of the principle of individual rationality. According to Hurley, individualism about agency is the view that the rationality of an act is determined by the causal consequences of that individual act only, rather than by the causal consequences of any collective act of which it may be a part (1989: 151). Hurley argues quite persuasively that individualism about agency is mistaken. In order to explain the rationality and the irrationality of certain acts we must acknowledge that the unit of agency is not fixed. Consider the case of voting. An individual may recognize that their own vote will not affect the outcome of an election nor will it cause others to vote. Even if it did have some minimal effect, the costs of voting would far outweigh these effects. How do we explain, then, why individuals should and do vote? The individualist about agency would have to argue that it is irrational for individuals to vote. Hurley points out, however, that if we allow that the unit of agency can be either an individual or a group then the rationality of voting becomes apparent:

> One may still vote because one's vote is partly constitutive of a valuable form of collective agency – valuable in terms of its causal consequences; it may be rational to, and irrational not to, co-operate with whoever is co-operating in this form of collective agency. In this case, one may act despite clearly recognizing the absence of certain causal consequences of one's individual act, in order to participate in a valuable form of agency. (1989: 149)

Although McMahon provides an extremely clear discussion, it is not clear to me that his solution to the problem of cooperation is not just

a restatement of Hurley's point. To be fair, he does mention Hurley in a footnote and suggests that the difference is that his account is more detailed. This is true. But then it appears that he hasn't offered us a new approach, just a clearer one.

McMahon also admits that the PCR could be modeled by standard rational choice theories (29). Cooperatively disposed people, on McMahon's account, are to be understood as assigning the same payoff to defection as they assign to non-cooperation. In other words, we assign cooperation the same payoff it would receive if everyone also defected. Thus, McMahon's account of rational cooperation involves a sort of universalization. A person is asked to consider whether one would endorse the performance of a certain action by everyone in the group. But the procedure of universalization can be modeled within standard utility theory by appealing to second-order preferences. We assume that people who recognize the requirement of universalization are rational agents who have a second-order preference to assign utilities on the basis of the procedure of universalization. "And similarly, we can suppose that agents guided by the PCR are standardly rational agents who have a decisive second-order preference to assign utilities to action in the way characteristic of that principle" (29). McMahon insists, however, that the PCR better elucidates the nature of rational cooperation. I'm not convinced. This isn't to say that McMahon's discussion isn't valuable. It is by far the clearest discussion of these issues I have encountered and this, in itself, recommends the book. But there is a lingering question about this particular aspect of the book – what is new here?

There is, however, something relatively new in subsequent chapters – the acknowledgment of collective agents. In chapter 2, McMahon explores cooperative schemes and the ways in which choice of a cooperative scheme is rational. Prisoner's dilemma cases present one cooperative scheme as salient but in many cases there will be several different cooperative schemes and these schemes will all be feasible according to the PCR. The members of a group of cooperatively disposed people can regard the choice of a cooperative scheme as guided by reason if they can regard it as justified by the principles or values, such as fairness, that they accept as governing their choices. According to McMahon, "when the cooperative potential created by general acceptance of the PCR within the group is actualized by the designation of a scheme, the group constitutes a collective agent" (40). McMahon provides the following example. A person falls through the ice in a pond. A bystander offers other bystanders a plan to save the person and this plan is enough to "crystallize" the group's cooperative potential. Collective agents can also be brought into existence when a group of agents establish a choice mechanism for identifying and implementing schemes when the opportunity arises – a voting procedure, for instance.

McMahon's discussion of collective agents is painfully underdeveloped and the discussion of collective intentions and collective wrongdoing that follows does not help clarify the nature of McMahon's claims. *Prima facie*, McMahon seems to be arguing for a form of collectivism – the view that groups have a unique ontological status. That is, they exist "over and above" their constituent members. Collective agents, according to McMahon, "come into existence" and "go out of existence" (40). He also claims that the state is a paradigm case of collective agency. "In applying these ideas to the political case, we can view the state as the cooperative enterprise and government as the mechanism that selects cooperative schemes by enacting laws underwriting them. This means that the state is a collective agent. This fits well with the idea, common since the seventeenth century, that states are artificial persons" (63).

To attribute personhood to a state, even artificial personhood, is to make a substantial metaphysical and ontological claim, but we find absolutely no defense of this claim other than a stipulation of when collective agents "come into existence." Are collective agents like human agents? What are the identity conditions for these agents? Do they survive changes in membership? Do they *really* have intentions or are our ascriptions of intentions to groups just shorthand ways of referring to the attitudes had by group members? McMahon fails to address any of these questions.

Part of the problem is that McMahon does not provide a context for his views. In the past few decades analytic philosophy has turned its attention to the issue of collective agency. Most notably, Margaret Gilbert (1988, 1994, 2000) argues for the view that individuals form plural subjects and that these plural subjects are the appropriate subjects of attitudes like belief and intention. Although McMahon mentions Gilbert and others in passing, there is no attempt to situate his view within social ontology. His discussions of collective intentionality and collective wrongdoing leave untouched the crucial question of whether there *is* a distinct collective subject that can be the bearer of intentions and responsibility. Thus, his discussion in this section of the book leaves one rather unsatisfied.

Even if we did get a fuller discussion and defense of collective agency, it appears to play absolutely no role in the rest of the book. To acknowledge that groups, rather than just the individuals that comprise them, can be agents would seem to open the door to the possibility that collective deliberation is a form of distributed cognition – cognition which occurs not merely in individuals but across the group. Collective reasoning on this view isn't just a way for individuals to arrive at justified beliefs. It is a process by which the group itself arrives at justified belief. Oddly, McMahon takes a much more individualistic approach to collective reasoning (chapter 5, e.g. 105–15).

Like individual reasoning, collective reasoning aims at justified belief or justified judgment. McMahon points out that there are two ways of understanding the cooperation involved in collective reasoning. On the first, the cooperation involved in collective reasoning involves the pooling of reasons. Individuals bring various reasons to the table and offer them for consideration by others. Individuals then make individual judgments on the basis of this pool of reasons. The second way of understanding collective reasoning is to see the collective process as extending to the judgment phase. After evaluating the reasons brought forth, the collective engages in further deliberation concerning what the reasons support and then it arrives at a joint decision. On this way of understanding collective reasoning the appropriate outcome of collective reasoning is not an individual judgment but a collective judgment. There are two ways of understanding this collective judgment. The collective judgment could represent a piecemeal consensus. Individuals, after reflecting on the relevant reasons, all arrive at the same judgment. The alternative way of understanding collective judgment is to view the process as resulting in an *integral* consensus. In an integral consensus, individuals suspend their individual judgments until some consensus is reached.

McMahon argues that collective reasoning ought to be understood only as a pooling of reasons. Collective reasoning, according to McMahon, is a form of mutually beneficial cooperative behavior. Each individual must benefit from the process of collective reasoning and the benefit they receive is the attainment of better justified individual judgments. Collective reasoning that results in an integral consensus, however, does not better justify individual judgments. And even if reaching a collective judgment is more beneficial the benefits could be derived equally well by reaching a piecemeal consensus rather than an integral consensus. As McMahon puts it, "the contact with the rational import of a given pool of reasons that an individual attains by participating in a collective judgment cannot be superior to the contact that he could attain independently. A collective judgment is simply a number of individual judgments packaged in a certain way. Thus, the contact with the pool that each makes by participating in a collective judgment is still contact that he makes as an individual. And this means that he could, in principle, have made it alone" (113). Understanding collective reasoning in this manner explains why individuals benefit from collective reasoning even when there is disagreement or a failure of consensus.

But understanding collective reasoning in this manner is to deny collective agency. The collective, itself, does not arrive at a judgment. What McMahon describes is best thought of as a form of *shared* reasoning. Individuals share reasons and arrive at individual judgments. Of course, this should not be troubling to the individualist who denies that there are collective agents, but one wonders how McMahon's

individualistic approach to collective reasoning meshes with his explicit acknowledgement of collective agents in an earlier part of the book (chapter 2: 34–40).

There is the further issue of whether we ought to buy McMahon's individualistic approach to collective reasoning. McMahon argues that we ought to understand the aim of collective reasoning as producing a pool of reasons from which individuals can draw to arrive at justified individual judgments. This is a normative claim. But collective reasoning often does result in an integral consensus. Is McMahon arguing that collective reasoning that aims at an integral consensus is irrational? In a pluralistic society we cannot always rely on the possibility of a piecemeal consensus. And there are good practical reasons for attempting to achieve an integral consensus – peace, for one.

We might also question McMahon's assumption that the benefit of collective reasoning is always a benefit to individuals. Collective reasoning is, according to McMahon, mutually beneficial cooperation. If individuals are not benefited then an individual has no reason to participate in the cooperative enterprise. But it may be that the benefit of collective reasoning is sometimes granted to the group rather than to the individuals in the group. Perhaps, as evolutionary biologists David Sloan Wilson (1989) suggests, there are certain traits, capacities, and processes that are beneficial to communities rather than any particular member of that community. Despite similarities to Hurley, McMahon seems guilty of *fixing the unit of agency*. If one continues to ask what is practically rational from the singular perspective of the individual, contributing to collective reasoning that results in an integral consensus will not be rational for the individual. If one asks, however, what is rational from the standpoint of the plural perspective, the answer may be that collective reasoning that results in an integral consensus is practically rational for the group.

Despite these concerns, McMahon's book is important and ought to be read. It makes a substantial contribution to discussions in political and moral philosophy and provides a clear and articulate discussion of central issues in rational choice and game theory. His solution to the problem of rational cooperation, if not novel, is more fully developed and defended than others. Although his discussion of collective reasoning remains individualistic, it is not unreflectively so. No doubt McMahon has a response to the anti-individualist approach I gesture at above. He presents a formidable challenge to those who want to argue that collective reasoning ought to result in an integral consensus.

We are not epistemic atoms. We engage in collective deliberation and reasoning on a daily basis and we engage in cooperative endeavors in virtually every aspect of our lives. A simple trip to the store might involve cooperation and collective deliberation between a husband and wife. This book is one of the few systematic discussions of this phenomenon and

it makes a substantial contribution to our understanding of ourselves as social agents.

Deborah Tollefsen

University of Memphis

REFERENCES

Gilbert, M. 1989. *On Social Facts.* New York, NY: Routledge.

Gilbert, M. 1994. Remarks on collective belief. In *Socializing Epistemology: The Social Dimension of Knowledge,* ed. F. Schmitt. Lanham, MD: Rowman and Littlefield, 235–6.

Gilbert, M 2000. *Sociality and Responsibility: New Essays on Plural Subject Theory*. Lanham, MD: Rowman and Littlefield.

Hurley, S. 1989. *Natural Reasons.* Oxford University Press.

Wilson, D. S. 1989. Levels of selection: an alternative to individualism in biology and the human sciences. *Social Networks* 11:257–72.