Economics and Philosophy, **22** (2006) 159–177

REVIEWS

doi:10.1017/S0266267106210824

Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory, Nicola Giocoli, Edward Elgar, 2003, x + 464 pages.

The fame of *Modeling Rational Agents* precedes it. Nicola Giocoli's book won the Best Monograph prize, 2004, from the European Society for the History of Economic Thought, and – in an earlier manifestation as his doctoral thesis – the Joseph Dorfman Best Dissertation Award, 2002, from the History of Economics Society. It does not disappoint the expectations thus aroused. Giocoli's account is powerful and fascinating, and elegantly presented.

Giocoli starts by suggesting that we can distinguish between the *body of knowledge* and the *image of knowledge* of a discipline, and that the two interpenetrate and interact. The *body* concerns the theoretical and empirical knowledge acquired by the discipline, as well as its methods and open questions. The *image* concerns what the discipline thinks it is and should be, how it presents and justifies itself, both to itself and to the world. The *image of knowledge*, the self-image of the discipline, will determine such matters as the open questions most urgently needing resolution; the grounds on which they are to be resolved; what constitutes an authoritative disciplinary pronouncement; and how novices to the discipline are to be inducted and socialised into it.

Giocoli's thesis is that over, very roughly, the century from the 1890s to the 1980s, a transformation took place in the *image* of neoclassical economics, accompanied by corresponding changes in its *body*. The image change that Giocoli identifies is from a 'system of forces' (SOF) to a 'system of relations' (SOR) view of what economics is about. The SOF image is the traditional view of the discipline as investigating economic processes, including equilibrating processes, generated by market and non-market forces. The SOR image presents economics as a discipline investigating the existence and properties of economic equilibria in terms of the mutual

consistency of the given formal conditions, and ignoring the processes required to generate and underpin it.

So – in terms of the open questions most urgently requiring resolution, the SOF view would point to the explanation of how and why a particular equilibrium is reached or operates as an attractor, while the SOR view would point to the logically possible, non-contradictory, existence of an imagined equilibrium position. The SOF image would suggest that a relevant argument would account for the influence of market and non-market forces, while for the SOR image, what is required is logical unassailability and economy of assumptions. Authoritative statements, according to the SOF view will be founded on the mathematics of classical mechanics, while for the SOR view techniques which favour consistency over calculability are fundamental. Finally, the appropriate university curriculum for the discipline is modelled on theoretical physics in the SOF image, and mathematics in the case of the SOR image.

This change in focus, Giocoli argues, is accompanied by a change in the discipline's understanding of rationality, from one that stresses optimising behaviour, to a more rarefied notion of logical consistency. The older view, for example, was based on the assumption that the behaviour of agents could be explained by the maximisation of a utility function; in the newer view 'utility maximisation' survives only in the sense that choice is conceived as being consistent with the agent maximising a utility function.

The focus of this story is a particular puzzle, namely the failure of neoclassical economics to adopt Game Theory, and, in particular, the concept of Nash equilibrium, in the immediate post-war period, given the present-day consensus that the Nash equilibrium embodies the discipline's most fundamental idea. The answer that Giocoli gives is that Game Theory and Nash equilibrium were ideas whose time had not yet come. Only once the transformation of the dominant self-image of economics from SOF to SOR had been completed, and the consistency approach to rationality had displaced the older view, was neoclassical economics ready to hear what the game theorists were saying.

The book consist of two parts, articulated by an 'interlude', and each consisting of two chapters, plus an introduction and conclusion. The introduction sets out the thesis just described, and then outlines mathematical formalism from Hilbert to Bourbaki, and the logical positivism of the Vienna Circle. These constitute the 'humus' in which two trends germinate – a trend within economics from Fisher, Pareto and Slutsky to Hicks, Allen and Samuelson, via Hayek, the Swedish school and Hutchison, and a mathematical trend exemplified by von Neumann and Morgenstern, and John Nash.

The first part of the book discusses neoclassical attempts to escape from psychology (chapter 2) and perfect foresight (chapter 3). These,

respectively, refer to the projects of freeing economic agents of any dependency on 'mental variables' and psychological processes, and of relaxing the classical requirement of perfect knowledge on the part of agents for the achievement of an intertemporal equilibrium of the system as a whole. Giocoli argues that the two projects were inconsistent and led to a stalemate lasting from the late 1930s until well after World War II, which was resolved only by the replacement of the SOF by the SOR approach as the dominant self-image of economics.

The second part of the book discusses von Neumann and Morgenstern's (chapter 4), and then Nash's (chapter 5), versions of Game Theory, and explores the puzzle, indicated above, of the fall and subsequent rise of Game Theory and Nash equilibrium in the post-war period. In a sub-plot to this account, Giocoli examines the writings of von Neumann, Morgenstern and Nash to see where they stood on the SOF-SOR issue – and finds an ambivalence with some strong evidence of a preference for the SOF version. This leads to the intriguing counterfactual speculation as to what might have happened had their contributions been sold vigorously to the profession as embodying an SOF vision. Neoclassical economics might, he suggests, have found an alternative resolution to the crisis of the SOF view which did not lead to a victory of SOR.

In the remainder of this review, I would like to do two things, both prompted by the observation that, remarkably, a book dealing with the evolution of the discipline of economics from the interwar period to the 1980s has nothing to say about Keynes. Firstly, the contrast between the SOF and SOR versions of equilibrium is not the whole story: there is a significant alternative to both that Giocoli ignores. Secondly, I wish to argue that Giocoli's account is also incomplete in another respect: he neglects macroeconomics, and once this is considered the resolution of the SOF-SOR rivalry looks different in important ways. These suggested extensions to *Modeling Rational Agents* are themselves a testament to the power and fruitfulness of Giocoli's approach.

I want now to turn to a consideration of a central theme of the book, namely, the neoclassical concept of equilibrium. Both SOR and SOF views embody notions of equilibrium, but, according to Giocoli, in the SOF image the focus is on 'the explanation of how and why a certain equilibrium has been reached', in contrast to the SOR image, the goal of which is the demonstration of the existence of an equilibrium, though (in Hutchison's words) 'not of [its] actual, empirical existence but of [its] conceivable, logically or mathematically non-contradictory "existence"' (p. 5).

Giocoli identifies the principal theme of the development of economics in the 1930s as:

> the last important attempt to preserve, if not enhance, the traditional image of economics as a discipline dealing with systems of forces, that

https://doi.org/10.1017/S0266267106210824 Published online by Cambridge University Press

> is, as a discipline which investigates the actual working of the economic system and, in particular, its equilibrating processes . . . [T]he key theoretical issues became the modeling of the disequilibrium processes . . . The program developed inside a more general theme, that of turning the static neoclassical equilibrium theory into a dynamic one. (pp. 135–36)

The SOF view was thus a view of economics as the study of the working of the economic system: the investigation of the equilibrating processes spontaneously invoked when the system was out of equilibrium. Giocoli argues that this attempt to draw dynamics from the static equilibrium theory was unsuccessful, partly because of 'unavoidable inconsistencies between the willingness to investigate the disequilibrium behavior of the economic system and the desire to preserve the notion of equilibrium as the central category of the analysis' (p. 137).

Giocoli touches here on some of the key issues concerning the way the equilibrium concept has been deployed in the neoclassical mainstream. Two things, I think, are clear from his account. Firstly, even the SOF version implies that the economic system can be understood as an equilibrium: the image of the economy as a whole is one of a static equilibrium, the maintenance of which is explained by the operation of equilibrating forces, forces which only operate once the equilibrium has been disturbed by exogenous forces. This leaves us with a profoundly static and ahistorical image of society: there is no theoretical basis here for immanent development or novelty. The recognition that the model might not be entirely adequate is addressed not by replacing it with an alternative, but by adding dynamics on to the static core, notably by relaxing the perfect information assumption and introducing various models of learning and expectations adjustment.

Secondly, the SOR version is clearly significantly worse, focusing the entire attention of the researchers involved on the study of theoretically conceivable equilibrium states, divorced from any possibility of learning about the equilibrating processes which might lead to and sustain such states. This, I submit, cuts us off from all possibility of learning about the forces which actually underpin and shape our society.

But there is an alternative to both. Throughout the history of modern economics there has been struggle, sometimes open, sometimes hidden, between two notions of equilibrium.

In the neoclassical view, an economic system is at or near a normal state or condition such that small moves away from it set in motion forces returning the system to the attractor state. The system can be modelled as an equilibrium. For some purposes, the equilibrium can simply be assumed to hold. If greater detail is required, a distinction can be made between a short and long run: in the long run, the system may be considered as, at least approximately, or for practical purposes, in the attractor state; in

the short run, firstly, changes in exogenous variables shock the system away from the attractor state, and then divergence of the system from the attractor state itself sets in motion forces returning it to its normal condition.

In the alternative to this view, self-organising economic systems exhibit stability underpinned by a host of adaptive mechanisms, that is, they are homeostatic. The terms homeostasis and equilibrium are often used interchangeably, but the notion of equilibrium here is fundamentally different from the neoclassical concept. In the 'years of high theory' prior to World War II, the neoclassical paradigm was challenged by the emergence from within itself of a number of standpoints which began to undermine the static notion of equilibrium. Two of the most notable challengers were Keynes and Hayek. In works such as the *Sensory Order* and *Law, Legislation and Liberty* Hayek refers with approval to the notions of homeostasis and open systems in the writings of Bertalanffy and others:

> In order to explain the economic aspects of large social systems, we have to account for the course of a flowing stream, constantly adapting itself as a whole to changes in circumstances . . . and not for a hypothetical state of equilibrium. (Hayek 1982: III, 159)

Keynes's *General Theory*, amongst other things, reintroduced many of Marx's insights, including the notion that the accumulation of capital over time must lead to a tendential fall in the rate of profit. Equilibrium here can only be a temporary halting place, an abstraction from the real flow, an assumption of *ceteris paribus* purely for analytical tractability.

What these paradigms are groping towards is a view in which time, change and history are fundamental, in opposition to the neoclassical view that equilibrium is the default, and disequilibrium a temporary disturbance which will spontaneously eliminate itself. In the former vision equilibrium is a temporary, short-run state in which growth and evolution are counterfactually assumed to be suspended. This temporary, provisional equilibrium is destined to be disrupted by the emergence of endogenous forces in the longer run. In the latter, neoclassical, view, stasis is the permanent, underlying condition, dynamics only arising temporarily after an exogenous shock, in the transition to a new equilibrium.

Giocoli's concern is to distinguish an earlier and a later phase of neoclassical economics and to account for the transition, while pointing up the deficiencies of the later SOR standpoint with respect to the earlier SOF view. My view, however, sketched out above, is that *both* the SOR and SOF images manifest a fundamental weakness intrinsic to the neoclassical outlook, namely the static neoclassical concept of equilibrium. The alternative, which was already emerging at the time, and which it was important to render heterodox, was the movement towards understanding the economy as a system – a *general* theory, as Keynes says, is 'concerned

with the behaviour of the economic system as a whole' (Keynes 1973: xxxii) – and, as Hayek argues, towards replacing equilibrium with homeostasis.

Turning now to the second of the two points I wished to raise, I'd like briefly to touch on the insights we might gain by supplementing Giocoli's account with a consideration of macroeconomics. By concentrating entirely on microeconomic fields such as general equilibrium theory, Giocoli's account misses the retention of the SOF image postwar within macroeconomics and econometrics. Following the aborted Keynesian revolution of the 1930s, academic economics and policy making circles were colonised by a bowdlerised neoclassical Keynesianism, and, subsequently, by an equally neoclassical monetarism. The period saw the radical separation of micro- and macro-economics (a separation denoted, paradoxically, as a 'neoclassical synthesis').

In Giocoli's account it is difficult to understand why SOR should have triumphed over SOF. It is clear from what he says that the SOF paradigm did not founder because it was eclipsed by SOR: on the contrary, although SOF was in trouble from the end of the 1930s, as the attempted escapes from psychology and perfect foresight produced opposing results – one leading away from, and the other towards, including agent learning – it was not until the 1980s that SOR was established as the core of the neoclassical standpoint, allowing the rediscovery of Game Theory and Nash equilibrium. So what drove this process? Why did it become untenable for economists to present themselves as following an SOF image, and why did they feel they had to switch to the SOR programme?

This makes more sense once we understand that, with the growth of neoclassical macroeconomics and econometrics, the SOF view was retained unchallenged in the postwar period. The SOF and SOR were complementary: in micro, including general equilibrium theory, there was an increasing recourse to maths and logic, and stress on the investigation of what it would mean for the individual agent to exhibit rationality, while in macro and econometrics, including mainstream Keynesianism and monetarism, there was little interest in the rationality or otherwise of the individual agent, since it was aggregate behaviours that were the focus of concern. Macroeconomics was interested in disequilibrium behaviour: how would agents respond to shocks while they were still learning about them? This was needed to underpin the development and use of fiscal and monetary policy for economic stabilisation. Microeconomics could be presented as attempting to supply rigorous foundations for this by explaining exactly what agent rationality meant, and what it would mean for the economy to be in equilibrium.

The two strands come together again with the emergence of the New Classical Macroeconomics (NCM) in the 1970s. Macro was then ready to adopt Game Theory, but by then Game Theory was no longer purely

SOR – as Binmore and others have pointed out. NCM attempted to build models which were internally consistent but which mimicked the progress of the economy. As Lucas makes clear in *Models of Business Cycles*, the focus of interest of the NCM was still the teasing out of dynamics from an equilibrium model, but now using the formalism of dynamic Game Theory. Although the NCM makes no attempt to discuss the behaviour of the economy in disequilibrium, it does attempt to make the model dynamic by incorporating exogenous shocks and then using Game Theory to model how agents will react to them, while invoking continuous Nash equilibrium conditions. It thus considerably expands the notion of equilibrium, calling on it to do much of the work done by disequilibrium in earlier schools of macroeconomic thought. It is thus problematic to fit the NCM into the pure SOR framework.

The strategy of orthodox, neoclassical economics was a reductionist one of basing everything on the individual optimising agent, set in a static social world, instead of on the evolution of systems of relations between agents, as the systems theoretical vision would have done. In that context the budding off of a microeconomics moving towards SOR, while leaving an SOF macro untouched, made explicit what was already implicit in pre-war SOF neoclassical economics: a need for a policy-relevant macro-level discipline, coupled with a putatively scientific account of the abstract agent on which it was founded. Much work needed to be done establishing the nature of this agent and the nature of the equilibrium constituting the minimal institutional framework within which these agents were to interact. The most stripped-down notion of the agent is one in which we know nothing of his desires, except that he pursues them consistently, and the most stripped-down social context is the Nash equilibrium in which everyone is behaving self-consistently, given self-consistent behaviour of everyone else.

I think this fascinating and provocative work can productively be linked to Mary Poovey's *History of the Modern Fact* (University of Chicago Press, 1998). The latter is an 'epistemological history' of political economy, a history of the ways political economists have sought to persuade themselves, and others, that what they are producing is reliable knowledge. Various phases of political economy are characterised, according to Poovey, by the metaphors and tropes, the rhetorical strategies which form, not only the language in which economics is communicated, but also the self-image of the discipline and hence the practice of economics itself. Without mentioning Poovey, this is precisely Giocoli's approach. SOF and SOR may both be understood rhetorically, as strategies to persuade us that economics is worth doing and the pronouncements of its practitioners worthy of attention. Both suffer from the perennial problem of the social sciences – the need to underpin their claim to the status of science, and both do so by reference to, and a claim to share the prestige of, a non-social

science – in the case of SOF this is classical mechanics with its associated differential calculus, for SOR mathematical formalism, combinatorics and set theory. To set this out explicitly raises interesting questions for further research, but to have set the scene for asking them is itself a very significant achievement.*

Andy Denis

City University London, and CPNSS, LSE

*An earlier and much shorter version of this review appeared in the *European Society for the History of Economic Thought Newsletter* 9, Summer 2004: 25–26.

REFERENCES

Hayek, F. A. 1982. *Law legislation and liberty*. Routledge
Keynes, M. 1973. *The general theory of employment, interest and money*. Macmillan
Poovey, M. 1998. *History of the modern fact*. University of Chicago Press

doi:10.1017/S0266267106220820

Microeconomics: Behavior, Institutions, and Evolution, Samuel Bowles, Princeton University Press and Russell Sage Foundation, 2004, 584 pages.

This important and highly impressive volume is intended as an overview of cutting-edge developments in microeconomics for graduate students. The book proclaims and upholds a strong 'institutional' and 'evolutionary' outlook, building on strong developments within economics in these areas within the last twenty years.

It is divided into five parts, consisting together of fourteen chapters. The first part is devoted to 'coordination and conflict' and consists of five chapters on institutional design, spontaneous order, preferences, coordination failures, cooperation and rent seeking. The second part addresses the institutions of capitalism, and covers topics including 'utopian capitalism', contracts, power, employment, wages, credit markets, and allocative inefficiency. In its three chapters, the third part discusses 'the coevolution of institutions and preferences'. The concluding part consists of a single chapter on 'economic governance: markets, states, and communities'.

The work is well written and carefully structured. Given its size and depth, detailed analysis and discussion would be impossible within the constraints of a short review. Here I must focus on a few key themes and make some relevant observations.

A principal thrust in its argument is to introduce matters of asymmetric information and power that were lacking in earlier mainstream accounts. This is demonstrated by comparing the discussion of 'utopian capitalism' in chapter 6 with 'the institutions of a capitalist economy' in chapter 10. The former capitalism is utopian because it unrealistically assumes complete contracting and unimpeded, efficient bargaining. Its free-market policy associations are thus based on shaky theoretical assumptions. The modern world, Bowles argues persuasively, is very different. In particular, employment contracts are both central to capitalism and by their nature incomplete in their specification. The final chapter gives an inspiring overview of the complex problems surrounding market and non-market economic organisation.

The *bête noire* of the volume is the 'Walrasian general equilibrium model' in which information is readily available, markets are complete, everything may be contracted, and there are decreasing returns to scale. For many, the escape from this Walrasian world was partly prompted by the negative 1970s results of Hugo Sonnenschein and others, who demonstrated that the excess demand functions in an exchange economy can take almost any form. Consequently, it became difficult to establish the uniqueness and stability results that were vital to the general Walrasian framework. The collapse of this Walrasian project was one of the major reasons for the mainstream turn to game theory in the 1980s. Another reason was the rise of institutional themes in economics since the 1970s. Once institutions enter the picture, their coordination solutions involve positive feedbacks that were absent in theoretical models based on pervasive diminishing returns.

Like other cutting edge theorists in mainstream economics, Bowles abandons the general equilibrium approach to make extensive use of game theory, and of theories involving incomplete contracting or asymmetric information. He is also often guided by results from experimental economics. To a large degree, therefore, Bowles's book reflects key mainstream developments that have become established in the 1980s and 1990s.

Nevertheless, his approach is distinctive, and does not follow every fad and fashion. For instance, he finds the concept of transaction costs highly ambiguous and definitionally slippery, especially where there is an interaction between contracting arrangements and technology (pp. 296–67).

The development of any academic discipline involves both continuities and discontinuities. The switch from Walrasian general equilibrium to game theory in the 1980s was an example of a discontinuity. However, many mainstream economists have retained a modified notion of individual rationality, expressing some conceptual continuity in the hard core of the science. Bowles (pp. 96–101) similarly defends a modified

concept of rationality from its critics, and is unhappy with the term 'bounded rationality' because for him it suggests a limit to rational capabilities.

Bowles is now situated in the phalanx of mainstream economic theorists. Yet his intellectual journey to this leading position was far from typical. In the 1970s he was primarily engaged in defending and developing aspects of Marxian economic theory. Bowles was never enamoured by Walrasian economics, and his adoption of game theory occurred for different reasons. He was influenced by the 'rational choice Marxism' or 'analytical Marxism' that emerged in the late 1970s, with the work of Jon Elster, John Roemer and others. These Marxists adopted analytical tools such as game theory to explain class conflict, inequality, and so on. In basing their analysis on a version of the individual rational agent, they drew fire from other Marxists. Ironically, 'rational choice Marxism' of the 1970s provided Bowles with a bridge towards mainstream economics, which he crossed in the 1980s or early 1990s. However, the former links with Marxism are evident in the book under review, with significant quotes and asides on the position of Marx.

Especially within a journal on the philosophy of economics, Bowles's current formulation and defence of the rationality concept is worthy of some discussion. Like many others, he rejects the narrower version of rationality in which individuals are motivated by their self-interest, without concern for outcomes experienced by others. Instead, for Bowles, individuals take account of the intentions and behaviours of others. They are rule-following and adaptive agents. Rules are supported by sanctions and internalised by actors, who learn to adapt to changing circumstances. Making a further and highly significant leap, Bowles accepts and argues that preferences are situationally specific and endogenous. Testimony from experimental economics and elsewhere suggests to Bowles that learning, experience and social interaction affect individual preferences. This takes rationality a long way from the earlier formulations of, say, Gary Becker and George Stigler. The word 'rationality' is retained but the meaning is expanded and significantly changed.

For Bowles, beliefs and preferences are foundational. This reflects a widespread tradition in philosophy and social science, about which pragmatist philosophers and others have misgivings. I will not elaborate on these here, other to point out that a key difficulty is to account for the evolution, development and physiological grounding of human capacities for belief and preference, without diminishing their meaning and significance.

Bowles sees beliefs as involving individuals' understandings of the consequences of actions. Preferences are described as reasons for behaviour. However, the word 'reasons' here becomes so broad that it involves habits, emotions, instincts and visceral reactions (p. 99). By

insisting that agents are rational, Bowles is saying that they act for reasons. But 'acting for reasons' becomes so capacious that it accommodates any form of impulse towards behaviour, deliberative or otherwise. It becomes a banal statement of the near obvious: that individual psychological impulses are behind all individual behaviours – that these behavioural events have a psychological cause. By this criterion, automata, bacteria or cockroaches might also be deemed rational. Continuity with the older mainstream tradition of individual rationality is preserved by removing from the concept much of its meaning.

Bowles considers the relationship between his broadened concept of rationality and the standard utility apparatus of mainstream theory. For him, utility maximisation is misleading when it comes to addictions and other dysfunctional aspects of human behaviour. Revealed preference theory describes behaviour rather than explaining it. The use of utility as an evaluative tool for economic outcomes should be separated from its use as an explanation of individual behaviour. Nevertheless, Bowles retains the utility function as a schematic model. For Bowles, when individuals 'act according to a complete and transitive utility function they are said to be *rational*' (p. 101). But note that the meaning of rationality in the quotation here is quite different in substance from his other notion that there are causal impulses (including reasons) behind behaviours. Bowles does not seem to notice this contradiction.

In both of Bowles's definitions, rationality is separated from another important connotation that it has occasionally acquired in economics and elsewhere. In much writing in economics, rationality takes on a richer connotation of deliberation, involving mental prefigurations and judgements. Such rationality involves 'reasons' in the narrower sense of codifiable and normative justifications for acting in a particular way. In broadening the concept of rationality, effectively to cover automata and amoeba, Bowles loses sight of such deliberative and prefigurative aspects of human thought and action. Simon was referring to rationality in a narrower and more deliberative sense when he coined the term 'bounded rationality'. Simon referred to limited deliberative and calculative abilities, and did not claim (against Bowles's broader sense of 'rationality') that action is sometimes without reason or cause. Accordingly, Bowles's criticism of Simon's concept somewhat misses its mark.

Although Bowles fudges the distinction between deliberative and non-deliberative behaviour, partly by describing both as 'rational', he sometimes insists on the importance of human intentionality, including in the processes of natural and cultural evolution. But the reader is explicitly forewarned (p. 60) that the nature and significance of intentionality is postponed to chapter 12. While the formal analysis therein (as in other chapters) is interesting and useful, little is added to the meaning of intentionality other than to contrast it with accidental or stochastic

behaviour. The formal models show that under specified conditions, groups of intentional agents among stochastic populations can jolt the system from one equilibrium to another. But this gives us little further insight concerning what human intentionality really means, or how it might contrast with non-random and goal-directed (group and individual) behaviour among other animals.

Two rhetorically striking features of this text are its insistence on the tenet of 'rationality' (albeit very broadly defined) and the conjoint and extensive use of various types of game theory. Strikingly, these two features are much less prominent in an alternative modern tradition of 'evolutionary economics' stimulated largely by the seminal work of Richard Nelson and Sidney Winter: *An Evolutionary Theory of Economic Change* (1982). Strangely, Nelson and Winter (1982) appears in the bibliography, but the names of these two authors are omitted from the index. Also unmentioned in the index are leading evolutionary economists such as Kenneth Boulding, Giovanni Dosi, Luigi Marengo, Stanley Metcalfe, and Ulrich Witt. Bowles proclaims the value of evolutionary theory for economics, but does not seem to want to converse with several important evolutionary economists. They were developing evolutionary themes some time before Bowles made evolutionary theory central to his own work.

Other glaring omissions exist: there are several mentions of Herbert Simon but none of the institutional theorist James March. Neither the word 'complexity' appears in the index, nor the name of Stuart Kauffman. Also omitted from the index are old institutionalists such John R. Commons, Wesley Mitchell and John Maurice Clark. Thorstein Veblen is mentioned in the text and index, but unfortunately mis-spelt four times. Forgotten too are *Marxisant* and other Cantabrigians who loomed larger in Bowles's earlier development, such as Nicholas Kaldor, Joan Robinson and Piero Sraffa.

Some errors are indicative of accidents rather than intellectual exclusions; for example, Sigmund (1998) is in the text but not the bibliography. Other absences are of unclear origin and explanation. For example, Bowles persuasively emphasises the frequent importance of communities, clans or ethnic ties in the enforcement of contracts, citing his recent joint work with Herbert Gintis, and the preceding works of Elinor Ostrom and William Ouchi, but not the pioneering and highly pertinent studies of Janet Landa, to which the later work of Bowles and Gintis in this area bears some similarity.

How does one explain these omissions? It cannot simply be that Bowles ignores work that he finds lacking in rigour or differing from his own approach. After all, he cites Simon despite his misgivings concerning the concept of bounded rationality, and he mentions Ronald Coase and Oliver

Williamson despite his well-grounded worries regarding the definitional vagueness of the concept of transaction costs.

I cannot explain these omissions here. But they are an additional source of interest to the reader, as well as the rich analytical content of the volume. Such questions would perhaps be answered by a rhetorical analysis, noting also Bowles's tactical broadening of the concept of rationality. I have little doubt of the rhetorical power, as well as the analytic force of this volume, and it will rightfully gain a place as a leading mainstream text, despite the references to Marx and the traces of Bowles's own distinctive intellectual evolution from Marxism.

Overall, this is a very fertile and inspiring book, of much broader use than its intended audience of graduate students. Its analytical accounts of institutional structures and its masterly fusion of institutional and evolutionary themes might eventually warrant its status as a modern classic. Yet at the same time it betrays an unwarranted narrowness of scope, where he is less embracing of accounts that would conflict with his (rhetorically significant but ultimately rather vacuous) account of rationality and his heavy reliance on game theory.

Geoffrey M. Hodgson

University of Hertfordshire

doi:10.1017/S0266267106230827

Welfare and Rational Care, by Stephen Darwall. Princeton University Press, 2002, xi + 135 pages.

A person's welfare is not what that person herself values, prefers, or wants – not even rationally. Rather, a person's welfare is what one ought to want for that person insofar as one cares for her. So claims Stephen Darwall in this short but engaging book. The concept of welfare, we are told, is inherently normative in the sense that it is conceptually linked to "an 'ought' or normative reasons claim" (4): a person's welfare just *is* "what one ought to desire and promote insofar as one cares for him" (7), or "what we would rationally desire for someone insofar as we care for her" (12). This metaethical theory of the concept of welfare – *the rational care theory* – is spelled out and defended in three of the book's four chapters. In the fourth and final chapter Darwall outlines and defends a substantive conception of welfare; what he calls *The Aristotelian Thesis*. This is the view that "the good life consists of excellent (meritorious or worthy) activity" (75).

The lead idea in Sidgwickian informed-desire accounts of welfare is that a person's good is the ideal object of that person's informed desires. Darwall criticizes such accounts for being "unacceptably broad [in that they] include within a person's welfare whatever he wants when fully informed" (27). We are invited to ponder the example of E.R. Burroughs's Tarzan. Tarzan and Jane are in love, but Jane is already promised to Clayton and she is determined to fulfil her duty and marry him. On reflection therefore, Tarzan prefers to leave Jane with Clayton. But as Darwall points out, Burroughs clearly does not want to suggest that this would be best *for* Tarzan, or *for* Jane, even though it accords with Tarzan's and Jane's informed desires.

A more general argument against informed-desire accounts is that "they make rational self-sacrifice conceptually or metaphysically impossible" (53). If Tarzan's good is equated with what he desires when fully informed, accepting the fact that Jane is determined to fulfil her duty would involve no self-sacrifice on Tarzan's behalf since that is what he on reflection prefers that she does. I am pretty certain that the Sidgwickian won't be persuaded by this argument, and that he probably shouldn't be. For why can't it simply be maintained that Tarzan's (and Jane's) acting is self-sacrificial insofar as they both on reflection prefer to obey the call of duty, even though they are aware that doing so is likely to cause them lifelong mutual regret? "Granted," the Sidgwickian could say, "obeying the call of duty enhances Tarzan's and Jane's good as this is what they on reflection prefer. But what makes this self-sacrificial is the simple fact that they deliberately abstain from what bears promise of being a much more joyous life."

Still, Darwall maintains that the upshot that obeying the call of duty enhances Tarzan's and Jane's welfare is embarrassing enough for the Sidgwickian account. This is one reason why we need a different conception of welfare – the *rational care theory*. This theory yields the right result since what we ought to want for Jane insofar as we care for her is that she sticks with her beloved Tarzan.

The rational care theory invites two immediate objections: (i) it appears to put the cart before the horse insofar as care seems to presuppose a notion of welfare – to care about someone is (partly, at least) to be concerned about that person's welfare; (ii) the theory seems terribly paternalistic insofar as it states that a person's welfare is wholly independent of what that person herself wants or would want after sober and informed reflection. Darwall anticipates and answers both of these objections. Beginning with (ii) he complements his rational care theory of welfare with a theory of respect (14–16, 43–49). The idea is that there is more than one way of valuing a person intrinsically. Even if *care* might prompt us to frustrate a person's wants and wishes, *respect* may tell us to do otherwise. We sometimes ought

to respect a person's considered and autonomous wishes even though we believe that it will be detrimental to her welfare. So Darwall can enjoy the attractions of a paternalistic theory of welfare without being forced to the disturbing upshot that we ought always to act paternalistically *vis-à-vis* other people. As for objection (i) Darwall is confident that it is 'care' and not 'welfare' that is the independent variable; it is the latter that should be defined in terms of the former (71). But how, then, are we to define 'care'? The bold suggestion that gets fleshed out in chapter 3 is that it need not be defined. Drawing on recent psychological literature on sympathy and empathy as well as on older philosophical writings on the same subject, Darwall suggests that we treat care as a "psychological natural kind" (50).

Interesting as this last claim is, I prefer to pay closer attention to the heart of the rational care theory, i.e., the metaethical claim that the concept of welfare is inherently normative. I will first raise some questions concerning the kind of normativity involved, and secondly, I will question the very idea that welfare is a normative concept.

Consider this crisp formulation of the rational care theory: "*What is for someone's good or welfare is what one ought to desire and promote insofar as one cares for him*" (7). I shall call the last part of this formula (the "*insofar as one cares for him*" bit) the "caring proviso." I get back to it shortly. Meantime, let us note that Darwall acknowledges that thus put, the normative status of welfare is akin to the normative status of the means/end principle that is so familiar from instrumental reasoning and which states that insofar as you desire an end you ought to desire the necessary means to that end. But having noted the kinship, Darwall is reluctant to say that "welfare's normativity [is] only hypothetical in the same way means/end reasoning is" (7). Welfare's normativity, then, is different. As Darwall points out, there are two ways of respecting the means/end principle of instrumental reasoning: either desire the means or give up the end. However, "caring for someone places one under a . . . consistency constraint of being guided by that person's welfare [where] the reasons [for being so guided] are not conditional on one's caring" (8). This strikes me as correct. Care does not seem to be – to borrow a piece of jargon from Derek Parfit – "conditional on its own persistence" (Parfit, *Reasons and Persons*. Oxford University Press 1984): I care about the members of my family and I want things to go well for them even if I were to deteriorate and lose my concern for them. But I fail to see how this is supposed to make the normativity of welfare any different from the normativity of the means/end principle. The fact that some of our desires are unconditional on their own persistence is a psychological fact about us. The mere fact that we desire certain ends, as it were, unconditionally does not credit these ends with a special normative status, for it is possible that some of these ends are very silly indeed. In

other words, the mere fact that some ends are desired unconditionally does not guarantee that there is "a 'categorical' reason for taking the means . . . to the relevant end" (7).

At times it seems that Darwall wants to say that there actually is categorical reason to care about people's welfare (but *cf*. 38). The depressed self-loather who fails to see why his welfare matters because he does not think that he is worth caring for is according to Darwall not *conceptually* mistaken. He is right in thinking that *if* he weren't worth caring for, his welfare would not matter, i.e. would not generate reasons. But, says Darwall, "it's just that he is wrong in thinking that he is unworthy of care. The deep truth that underlies the depressive's claim is that it is a person's being worthy of concern . . . that makes considerations of his welfare into reasons" (6). Darwall's Kantian leanings can be traced here. The picture he presents us with seems to be that persons have worth simply qua persons. And to say that a person has worth is to say that there is reason to care about that person (84).

This gives rise to two questions: First, it is no longer clear what role the "caring proviso" is supposed to play in the definition of welfare. Remember that we were told that a person's welfare is what we ought to want for that person *insofar as we care for him*. Now we are told that persons have worth simply qua being persons. And for a person to have worth is for that person to be such that there is reason to care for him. So why not simplify matters and drop the now otiose "caring proviso"? A person's welfare may then be defined simply as what one ought to want for that person (where the "ought" is – as before – to be understood in "its most general normative sense" (8)). Second, Darwall's broadly Kantian theory may be vulnerable to a standard critique of consequentialist theories, namely the charge of over-demandingness. If we assume, as seems reasonable, that all persons have equal worth we seem to end up with the view that we have reasons to care equally for all people. This does indeed seem demanding. There may be possible ways out of this predicament but Darwall has yet to tell us what they are.

Let us now turn to the deeper question whether welfare really is normative. It is instructive to compare the argumentative structure of *Welfare and Rational Care* to that of another important and fairly recent contribution to the field; Wayne Sumner's *Welfare, Happiness, and Ethics* (Clarendon Press 1996). While Darwall opens his book by claiming that welfare is a normative concept and closes it by arguing in favour of a certain substantive conception of welfare, Sumner proceeds in the almost completely reversed order. He is mainly concerned with arguing for and against certain substantive conceptions of welfare and ultimately settles for the view that welfare is to be understood, roughly, as "authentic happiness." Only in the final chapter does Sumner go on to argue that welfare so conceived ought to be promoted for its own sake (indeed that

it is the only thing that ought to be so promoted). Darwall and Sumner can thus be said to represent two rival camps in the metaethical debate on welfare: Darwall is a representative of the view that welfare is an inherently normative concept (a normative reason claim follows analytically from the claim that something would enhance someone's welfare); Sumner is a representative of the view that welfare is not inherently normative (a normative reason claim follows from the claim that something would enhance someone's welfare only given certain substantive views). Which one of the two views has the most to be said in its favor?

Darwall has two main arguments for his claim that welfare is inherently normative. The first goes as follows:

> [I]t seems possible for two people who care about someone, *S*, to coherently disagree about whether something, *X*, is good for *S*, even though they agree completely about all the non-normative facts concerning *X* and *S*. . . . Suppose for example that *X* is a pleasant illusory belief of *S*'s, say, that *S*'s novel has sold 10,000 copies (when in fact it has sold only 12). It would seem that two people could be agreed about everything else, but simply disagree about whether this pleasant illusory belief is good for *S* or makes some contribution to his welfare, other things being equal. In such a case, it is hard to see what else they could be disagreeing about other than whether *X* is to be (ought to be) desired for *S*'s sake, or, equivalently, whether it would be rational (warranted, justified, make sense) for someone who cared about *S* to desire *X* for *S*. (11)

A first thing to note is that it is question begging to assume at the outset that the two antagonists in this scenario do agree on all non-normative facts. Darwall's thought must be that they agree on everything regarding *X*, except for whether *X* contributes to *S*'s welfare. One explanation of what goes on here would be that the disagreement is normative, and that welfare consequently is a normative concept. This is the conclusion Darwall wants us to draw. But the argument is a *non sequitur* as the following analogous scenario illustrates: Rock'n'roll connoisseurs Beavis and Butthead have a disagreement. Beavis thinks that a certain album is a Heavy Metal album, while his companion Butthead thinks not. Being connoisseurs, they both care about getting it right, and furthermore, they agree about every aspect about the album except for whether or not it is a Heavy Metal album. Now, I am sure that no one would be tempted to draw the conclusion that the disagreement here is normative and that Heavy Metal is a normative concept. A sensible explanation of what goes on is that Heavy Metal is a "genre concept," and that standards for applying such concepts vary among people. The point of the analogy is that we may explain the disagreement between Darwall's antagonists in a similar way, i.e., not by appealing to a normative concept of welfare but by treating welfare as

a genre concept. So it does not follow from Darwall's scenario that welfare is normative.

Darwall might object that this prevents the antagonists from *coherently* disagreeing since they apparently mean different things by the term "welfare." But isn't it quite common in philosophical debates that philosophers disagree on everything regarding some *X*, except for whether *X* is *Y*? It is unlikely that in all such cases the disagreement, and consequently the concept *Y*, is normative. Perhaps such disagreements are too fundamental to be called 'coherent disagreements'. But then it is not clear after all that the antagonists in Darwall's scenario can coherently disagree.

The second argument that Darwall frequently appeals to is that the rational care theory neatly explains why reasons generated by welfare considerations are not merely agent-relative but agent-neutral, i.e. graspable "from the perspective of someone who cares for the person" (45, *cf*. 14–15, 25, 46–47, 72, 83, 98). But we need not build normativity into the concept of welfare in order to hold that welfare considerations generate agent-neutral reasons. We could equally well accept Sumner's non-normative conception of welfare as "authentic happiness," and then go on to argue that it ought to be promoted. This, I take it, is the standard welfarist utilitarian line.

Moreover, I believe there are at least two considerations that should make us hesitant to accept the idea that welfare is a normative concept. The first is that if we do, we run the risk of being unable to make conceptual sense of substantively crazy but coherent views about how to live. Imagine a religious sect of extreme ascetics whose members firmly believe that they ought to devote their entire lives to the compensation of sin. Assume that they care about each other and that each member sincerely believes that what he ought to want for himself and for his fellow sectarians are constant hardships and pain. Were we to follow Darwall, we would have to say that the ascetics' views about how to live are their view of welfare. Substantively crazy, of course, but a view of welfare nonetheless. My worry is that this does not take the ascetics seriously enough. What makes them so utterly different from us non-ascetics is not merely that they have different views about welfare, but that they shun welfare altogether. According to their view of how to live, welfare is not what you ought to want for those you care for. The rational care theory rules this out as incoherent. The second consideration that makes me hesitant to accept the idea that welfare should be seen as a normative concept is purely economical. We have a plethora of normative concepts already, some of which Darwall makes frequent use of in explicating the normativity of welfare (e.g., ought, reason, rationality, etc.). Why introduce yet another one if it seems that we can do without it?

In the fourth and final chapter Darwall claims that his metaethical theory of welfare, the rational care theory, and his favored substantive

theory, the Aristotelian Thesis, mutually support one another. It is not clear to me how this is supposed to work. It is not even clear to me that substantive implications are virtues of metaethical theories in the first place. And the fact that Darwall concedes that "questions of normative ethics are logically independent of metaethical issues, and this is no less true when it comes to welfare" (73) renders his train of thought here mysterious.

Although I don't quite see the connection between Darwall's metaethical theory of welfare and his substantive theory and may therefore have missed the ultimate twist in the argument, I warmly recommend the book. Darwall's prose is as elegant and captivating as ever, and anyone with an interest in welfare, metaethics, or moral psychology will find useful things. There are not many pages in *Welfare and Rational Care*, but there sure is a lot to think about.

Jonas Olson

University of Oxford

https://doi.org/10.1017/S0266267106210824 Published online by Cambridge University Press