Economics and Philosophy, **16** (2000), 147–174

REVIEWS

Individualism and the Unity of Science, Harold Kincaid. Rowman & Littlefield, 1997, vii + 165 pages.

The appeal of reductionism has waxed and waned over the years. During the heyday of positivism, it was quite the rage. Recently, however, reductionism has fallen on hard times. The modern allure of reductionism can probably be traced back to the triumph of Newtonian mechanics. Physics appeared as a master science. It revealed the power of the mathematical method in reducing the complexities of material nature to regular laws. Newtonian mechanics served as the paradigm of a proper science. The eighteenth and nineteenth centuries saw one attempt after another to formulate an appropriate 'Newtonian' science of psychology, chemistry, biology, geology, sociology, politics and economics. In addition, physics appeared to deal with the most fundamental phenomena of nature, so, in addition to serving as a methodological model, physics appeared to be the foundational science in a hierarchy. The appeal of this approach derives from the sense that, once a particular part of a complex system is reduced to order, there must be a simple set of rules and regularities in terms of which the whole becomes comprehensible. Once we have conceptualized nature as *one* system, we are driven to think that there must be *one* key to unlocking the secrets of that system. If the system exhibits hierarchical structure, as the natural and social world do, then if we can discover the laws of the fundamental units of the lowest level of the hierarchy, then we have the key to understanding the whole. Such are the intuitions that drive the reductionist program in all its manifestations.

In the social sciences, the reductionist wars take the form of battles between 'individualists' who argue that the key to understanding social phenomena lies in appealing to the properties and behaviors of individual agents and 'holists' who argue that some, if not all, social phenomena are irreducible. Harold Kincaid's *Individualism and the Unity of Science* is a subtle and nuanced analysis of the interlocking themes and

issues surrounding this struggle. Two major claims, one substantial and one methodological, emerge from this analysis. The substantial claim is a defense of a 'non-reductive unity' of the sciences. The 'nonreductive unity' favored by Kincaid embraces four basic tenets: (1) the explanatory power of the special sciences cannot be reduced to that of some fundamental or basic science; (2) explanatory independence is compatible with ontological dependence; (3) scientific unity is achieved through 'integrative testing' of the special sciences against lower level counterparts, etc., so, in a pluralistic unity, there is room for both molecular biology and biochemistry, for sociology and psychology, and so on; (4) the unifying glue that holds the package together is provided by 'interlevel theories' that connect the *sui generis* components from different levels of organization (p. 6). The key methodological point is that the disputes between reductionists and pluralists or between individualists and holists are *empirical* and not *conceptual* disputes.

There are more interesting themes in Kincaid's book than can be fruitfully explored within the confines of a compact review. Here I will focus on what I take to be the central theses: (1) the dispute between individualists (reductionists) and holists (anti-reductionists) is an empirical not a conceptual dispute; (2) approaches that are typically taken to be exemplars of individualism often, in fact, employ non-eliminable holistic assumptions.

1. THE EMPIRICAL CHARACTER OF THE DEBATE BETWEEN INDIVIDUALISM AND HOLISM

Kincaid's fundamental methodological thesis is the claim that whether a given phenomenon can be wholly understood in terms of the actions of individuals is not something that can be decided by appeal to general criteria. Each case must be decided on its own merits. Kincaid's own pluralism attempts to steer a middle course between the two extremes. The issue of monism versus pluralism can be put in the following way. There are three fundamental alternatives for describing and explaining the social: we can (1) opt for Individualism, I; (2) opt for Holism, H; or, (3) opt for Pluralism, which involves some combination of I + H, where 'I' and 'H' can be taken to stand for individualistic and holistic factors respectively. The reductionist question is: can the H factors be completely eliminated in favor of the I factors? Kincaid argues that, in general, they cannot. He opts for pluralism.

Kincaid takes as his primary foil the individualist thesis that 'we can understand everything we want to know about the social world entirely in terms of the actions of individuals' (p. 1). Kincaid argues that this thesis is false. In the social sciences, 'individualism … frequently fails because it presupposes rather than eliminates background social struc-

https://doi.org/10.1017/S0266267100240182 Published online by Cambridge University Press

ture' (p. 7). In order to assess this claim, we need to get a handle on exactly what the thesis of 'individualism' amounts to. Unfortunately, as Kincaid's analysis brilliantly shows, this is no easy task. He distinguishes four generic forms of 'individualism', some of which come in alternative versions.

First, there is ontological individualism: This doctrine comes in two forms: OI-1: social structures do not *exist* separately from individuals; OI-2: social structures do not *act independently* of individuals (pp. 13f.). Second, there is theoretical individualism: TI: 'all social explanations can be reduced to theories about individuals' (p. 14). Third, there is explanatory individualism: this principle has three versions: ExI-1: 'full explanation requires reference solely to individuals'; ExI-2: 'full explanation requires *some* reference to individuals'; ExI-3: 'purely individualist theories suffice to fully explain' (p. 14). This last version, ExI-3, is compatible with the thesis that holist accounts *can* be explanatory as well, but ExI-1 and ExI-2 rule this out! Finally, there is evidential individualism: this principle has two versions: EvI-1: 'all evidence is 'evidence about' individuals in some sense'; EvI-2: 'no social accounts is well confirmed until we have evidence about individuals, particularly individualistic mechanisms' (p. 14). These views are related to one another in complex ways and Kincaid's subtle analysis shows that those who seek to wade into the fray need to be more careful and sensitive than most have been in the past.

John Watkins (1973) characterizes individualism in terms of two theses: W-1: the ultimate constituents of the social world are individuals; W-2: social events are brought about by people. Kincaid characterizes W-1 as an 'exhaustion' principle and W-2 as a 'determination' principle. Their force, he argues, is that the social *supervenes* on the individual. So, one can endorse a form of 'individualism' without thereby committing oneself to reductionism. Just how plausible are W-1 and W-2, Kincaid asks? He claims that both W-1 and W-2 are *empirical* guesses about what the best theories now or in the future will do. What, exactly, theories and explanations in the social sciences are, or should be, designed to do are issues addressed more fully in Kincaid (1996).

There are three reasons to think that explanatory reductions, in general, will fail . First, social events and processes are likely to be multiply realizable. Second, individual actions have alternative descriptions depending on context. Third, 'any workable individualist social theory will in all likelihood presuppose social facts' (p. 33).

In effect, the first and second reasons are a reflection of the fact that maps from the individual level (I) to the social level (S) are many-many. Multiple realizability amounts to the claim that the maps R: S → I are 1-many; Alternative descriptivity means that the maps H: I → S are 1-many as well. If reduction requires lawlike co-extensionality between

predicates at the individual and social level, these reasons are telling. If not, then they are not. Kincaid considers and rejects some attempts to defuse the anti-reductionist force of the first two reasons. Some have argued that multiple realizability can be overcome by using disjunctive clauses to effect the reduction. So, if holist property h can be realized as i_1 or i_2, one can 'artificially' create a disjunctive property i_3 (= 'i_1 or i_2') and the multiple realizability apparently disappears. The problem, Kincaid notes, is that such ploys provide only 'accidental' co-extensionality Kincaid takes reductionism to require (pp. 38f.). Rosenberg (1994) makes a similar point in connection with the alleged reducibility of genetics to molecular biology. A second proposal is to reject the requirement that reductions require biconditional connections between T_h and T_i. If this *were not* a requirement, then multiple realizability would not pose a problem. Kincaid does not offer a knockdown argument against this, but he rejects a proposal along these lines offered by Mellor (1982). Mellor argues that the reducing relationships can be approximations and that such approximations are readily available for the various special sciences. Kincaid rejects both claims. A third proposal is to argue that reduction does not require any bridge laws at all. Kincaid rejects this as a conflation of theoretical reducibility and explanatory reducibility (p. 39). So, as far as he can see, and I take his case to be a strong one, the first two reasons for having doubts about the viability of purely individualistic explanations of social phenomena are telling.

The third reason for having doubts rested on the allegation that so-called individualist accounts often rest on or presuppose social facts. For example, some theories about the behavior of individuals appeal to preferences. But where do these preferences come from? Are they innate or do they have social explanations? Of course, for limited purposes, we can ignore these questions and treat the preferences as given. However, in the individualism–holism debate such questions cannot be ignored because how we answer them determines whether we have 'full explanations' at the lower level or not. If preferences do have social explanations, then they are effects that are functions of social contexts. This, in turn, suggests that social information is crucial to the implementation and application of what are supposedly 'fully individualistic' theories. Now this raises another question about completeness or full explanation. In one sense, no explanation or set of explanations, no matter how comprehensive, can be full or complete. All explanations, to appropriate a remark by Wittgenstein, come to an end at some point in the given. This point may be a 'floating point' but it cannot be overcome by appealing to further explanations without introducing a new stopping point in its stead. So, all theories are going to be saddled with concepts that they employ but cannot further analyze or explain. In the present context, the question is whether or not the best 'final theory' will or will

not include essential reference to irreducibly social factors. Kincaid's claim is that this is an empirical question, one that cannot be decided a priori on conceptual grounds. Three examples give this worry some bite.

Consider, for example, the theory of 'organizational ecology'. It uses population genetical models of organizations and the competition for resources. This approach introduces natural selection accounts. But, natural selection models are 'multiply realizable'. So, no reduction to purely individualist level is achieved by these approaches. In general, Kincaid concludes, 'social theories will be irreducible when they describe selection processes at the institutional level' (pp. 19f.). Consider, as another example, economic theories of the firm (p. 20). If firms are treated as black boxes (as individual preferences were in the first example) we have another example of a concept – at the social level – whose ultimate reducibility to the purely individual is a matter of faith. Whether it is achieved or not is an empirical question and not a result forced by any conceptual analysis. Finally, consider rational choice theory (p. 20). Gary Becker's account rests on individual preferences for social goods that seem to call for holistic or social explanations (1976, 1981). Rational choice theory, far from reducing social facts to individual facts, presupposes them. In general, 'rational choice accounts ... generally rely on a background of social institutions and process'.

The main point is that whether such reductions will succeed or not is an empirical question to be decided on a case by case basis. On the surface, this result might be seen to threaten the unity of the sciences by leaving it open whether any general reduction program will work. But, once we accept the possibility of science as a 'non-reductive unity', this piecemeal approach does not threaten to lead to a disunity of the sciences (cf. p. 66).

2. THE NON-REDUCIBLE SUPERVENIENCE OF THE SOCIAL ON THE INDIVIDUAL

Kincaid argues that social facts supervene on individual facts but are not reducible to them. It follows that purely lower level accounts are incomplete (p. 70). The gist of the problem is that supervenience, by itself, only guarantees that one kind of fact is 'fixed' by other kinds of fact. But, reducibility requires a number of other constraints as well. Whether these other constraints are satisfied in particular cases is an empirical question that needs to be settled on a case by case basis (p. 74). Jaegwon Kim's arguments are the foil here (Kim, 1993). How, Kim asks, can higher level structures exhibit any 'real' causal efficacy if, indeed, they are composed of and 'fixed' by lower level structures and processes? Kincaid's defense rests on an appeal to natural kinds. Without delving into the mysteries of what we should understand these

kinds to be, we can attest that they serve as the 'categories that play a central role in explanation' (p. 75). So, if higher level structures are genuinely causal, there must be higher level 'natural kinds'. If lower level structures are genuinely causal as well, there must also be lower level 'natural kinds'. Whether the higher level causal relations are epiphenomenal or not turns on whether the higher level kinds can be 'reduced' to the lower level kinds. But, in the light of the triumvirate of reasons for thinking that reductions, in general, will fail, in particular, in the light of multi-realizability, we can grant that higher level structures are *token* identical with lower level structures without assuming that *higher order kinds* are identifiable with *lower level kinds*. They may or may not be, but a *general* acceptance of supervenience does not commit us to a general acceptance of *type-type* identifications. If higher order kinds are not reducible to lower level kinds then the fact that they supervene on lower level tokens does not establish that higher level causal links are impotent.

There are two opposing views. (1) Danto (1973), Rosenberg (1985), and Stich (1985) all argue that individualist explanations are best and fully explanatory *even though* reduction is not feasible. (2) Dennett (1969), Garfinkel (1981), Putnam (1981) and Wimsatt (1976) argue for the opposite tack: lower level supervenient structures explain nothing about the higher level structures they constitute. If either of these claims were true, they would threaten Kincaid's picture of a non-reductive unity. To show that these claims can be defused, Kincaid appeals to the pragmatic conception of explanation. On this view, an explanation is a polyadic relationship which takes into account not only the explanans and explanandum but the interests of the inquirer and the context in which the call for an explanation arises. Garfinkel (1981), and van Fraassen (1980) present versions of this model. With this conception in hand, one can see how supervenient explanations may, on the one hand, provide adequate answers to *some* questions – that is, those that deal with the components of higher level structures, without being able, on the other hand, to answer *all* questions about higher level structures. This will be true if (as we assume) there are relevant higher level *kinds* which cannot be 'reduced' to lower level kinds (p. 81).

Kincaid's defense of his pluralistic unity rests on two key assumptions: (1) that there are 'important' questions that higher level theories can answer but lower level theories cannot, and (2) that higher level theories are genuinely explanatory (p. 82). As examples Kincaid cites the examples of understanding persons (which seem to rely on irreducible psychological concepts such as 'rationality') and our understanding of ideologies (which appeal to social structures and evolutionary theory). Prima facie, understanding persons and ideologies would seem to qualify as 'important questions' that, if Kincaid is right, cannot be

adequately dealt with by lower level theories. Indeed, if 'context sensitivity' and 'reliance on higher order presuppositions' is granted, then lower level theories make non-eliminable reference to these structures.

The arguments by Putnam and Garfinkel to the effect that lower level theories do not explain at all, appeal to the fact that micro-level explanations offer irrelevant details that macro-level accounts can ignore. Thus, in Putnam's example of why a square peg does or does not fit into a round hole, appeal to the micro-structure of the objects involved seems unnecessary. Kincaid objects, however, that this assumes that the micro-level accounts are supposed to provide micro-level types to match the macro-level types that, in this case, seem to doing all the explanatory work – namely, circularity and squareness. Of course, as we have just seen, multiple realizability precludes this. But, Putnam and Garfinkel are too precipitous in denying any explanatory power to the lower level accounts. If we look at the tokens and not the types, then the micro-level accounts are (or can be) explanations.

Since, on Kincaid's view, questions of reducibility and explanatory power have to be dealt with on a case-by-case basis, no general argument is forthcoming. As an illustration, Kincaid considers the case of neoclassical economics (Chapter 6). Neoclassical economics, despite its shortcomings, is often defended as providing the 'best explanation' of a wide variety of economic phenomena. Kincaid argues that the inadequacies are not outweighed by the explanatory power (p. 91). Explanations appealing to neoclassical economics rest on appeals to inferences to the best explanation. Inference to the best explanation (IBE) is often claimed to be a 'foundational principle'. Kincaid rejects that characterization and he rejects the claim that neoclassical economic explanations *always* provide the best explanation of economic phenomena.

By a 'foundational principle', Kincaid means that: (1) the principle in question must be a primitive strategy, that is, an unjustified justifier; (2) the principle must be purely formal; and (3) the principle must be sufficient, that is, given the data and the competing hypotheses, the outcome of the application of the principle must be indefeasible (the outcome cannot be overridden by appeal to other principles (p. 95)). But, Kincaid argues, inferences to the best explanation 'rest on substantive, contingent, and often implicit assumptions to do their work' (p. 92). As such, they cannot be foundational in the required sense. In order to assess this claim, we need some view about what constitutes 'explanatory power'. This is a contentious issue in its own right. Kincaid considers two accounts: (1) unification (à la Kitcher) and (2) the 'ability to cite causes'. Both, Kincaid argues, appeal to empirical, substantive claims.

The argument he gives is as follows (pp. 97f.):

1. Explanatory power is cashed out either in terms of the power to unify or the power to cite causes.
2. IBE as inference to the most unifying theory is just the principle: choose the theory that is best confirmed.
3. Therefore, IBE as IBUE (Inference to the Best Unifying Explanation) is *not* a special principle at all.
4. IBE as IBCE (Inference to the best Causal Explanation) is neither formal not sufficient.
5. Therefore, IBE as IBCE is not foundational.

The crucial assumptions are 2 and 4 and Kincaid provides arguments for them. The argument for premise 2 goes like this: IBUE can be understood in a number of ways. First, what does unification amount to? If it is cashed out in other epistemic terms (Harman, 1965: less *ad hoc*, more plausible; Howson and Urbach, 1993: best supported by the data) then it is not primitive. In fact, Kincaid suggests, on any such reading the concept is empty. This sounds right to me. I have always wondered what the big deal about IBE is. Kincaid's analysis suggests that, on some standard accounts, it is not doing much of anything.

If unification means cohering with a set of beliefs or 'fitting' with the 'most comprehensive argument strategy' (as Kitcher, 1989 suggests), then IBE is not trivial but it is neither formal nor sufficient (p. 98). IBE would be defeasibile on such a reading because if a hypothesis is an IBE if and only if it best coheres with a set of beliefs B we have about some empirical system, then we need to presume that the empirical system we are trying to account for is, in fact, best characterized by B.

The example Kincaid suggests is this. We have a set of basic beliefs about Darwinian selection systems, which we can label 'D's'. We come across a population P whose dynamics we want to account for. Suppose there are two competing hypotheses, NS (Natural Selection) and RD (Random Drift). The IBE that singles out NS on the grounds that NS best coheres with D will only be the best explanation of the dynamics of P if, in fact, the population P is adequately characterized as a D-system. Of course, we can expand what it means to be a D-system by including alternative mechanisms such as random drift but this does not affect Kincaid's point. The point is that the principle of choosing between H_1 and H_2 (where these are competing hypotheses to explain the behavior of some system S) on the basis of IBUE depends for its application in particular instances on *empirical* assumptions about the system S. IBUE, so construed, is neither formal nor non-defeasible.

It is certainly not formal unless the characterizations of the system S are taken to be part of the *data*, in which case, applying IBUE to a system S under one description might very well yield different results from applying IBUE to the same system under a different description. Given such a case, one might argue that what we have shown is not that the

principle IBUE is defeasible but rather that our descriptions or characterizations of systems are defeasible. Consider, for example, *modus ponens*, or MP. Surely, MP is a formal principle if any is. But, applying Kincaid's test, it might seem not to be. Suppose someone reasons as follows,' 'If Tully is a Greek, then Tully is mortal. Tully is a Greek. Therefore, Tully is mortal'. We then point out that Tully is not a Greek but a Roman. On this re-interpretation of the empirical data, the inference is not justified. Well, what do we say? We do not conclude that MP is not a formal principle. We say, rather, that the argument as originally presented was unsound. So it appears that the *application* of formal logical principles to particular cases can go astray without our concluding that the inference principles employed are themselves *non-formal*. What is supposed to be different about the IBUE case?

Consider the construal of IBE as IBCE. Kincaid argues that so construed, its applicability hinges on assumptions about the nature of causation and assumptions about what causal variables are or are not relevant in a given case (p. 99). These are clearly substantive assumptions and if they are presumed *not* to be part of the data to be explained, then the application of IBCE rests on substantive assumptions and is not a purely formal principle either. But, the same worry about the non-formality of IBUE can be run through here as well.

What are we to say about these cases? One might argue that the counter-example appeal to the principles of logic is not legitimate on the grounds that logical principles are *not primitive* in Kincaid's sense. Thus, we do not accept MP as a logical principle *sui generis*, but because it is truth preserving, etc. Reflection on this point brings one to the edge of the abyss of the justification of deduction and we will go no further. But, it does point out that even in the most formal of so-called formal reasoning, it is sometimes a struggle to separate out what is empirical assumption from what is 'formal' principle. Second, we can shed some light on the matter perhaps by recalling the claim that explanation involves the application of a model to an empirical system. What the implications of the model assumptions are, under some principle of explanatory inference, are one thing; whether a given empirical system does or does not conform to these assumptions is another.

Kincaid illustrates what he sees as the problems for both IBUE and IBCE through the case of neoclassical economics (pp. 99f.). One might object that 'neoclassical economics' is not so much a specific theory as an approach. To make his case, Kincaid proposes to take neoclassical economics to comprise eight claims (p. 93). (1) 'Economic outcomes must be explained as entirely the result of individual choices'; (2) 'Those choices are rational'; (3) 'Rational choices are those that maximize self-interest given constraints'; (4) 'Choices are coordinated by markets'; (5) 'Markets are best understood by focusing on full competition and

equilibrium outcomes'; (6) 'Full competition entails complete prices, full information about prices and technology, price-taking behavior by firms and consumers, and free flow of resources to new uses'; (7) 'Markets produce efficient outcomes – firms equate marginal revenues to marginal products and so on'; and (8) 'Incomes are returns to factors'.

It is often argued that (1) neoclassical theory is an individualist approach, and (2) neoclassical individualism is *incompatible* with non-individualist alternatives. Kincaid argues that both are suspect. As for (1), Kincaid notes that neoclassical theories invoke the concept of 'firms' – a social entity – (often) as an unanalyzable primitive. Households, representative agents, etc., are likewise all social terms that are not reducible to 'individuals'. So, even if neoclassical theory were successful, it would not provide unambiguous support for individualism. As for (2), although neoclassical accounts, on the views of some philosophers, compete with alternative holist accounts, there are other philosophers who see the accounts as complementary. Some aspects of our social life are either presupposed by neoclassical accounts or are beyond (or outside) their scope. Often, the neoclassical accounts are held to complement other accounts. On this pluralistic view, given the richness of our social life, there is enough for everyone to do. If the accounts are not competing, then no appeal to the IBE has been invoked – at least, no general appeal. It still may be the case that some aspects of our social life are explained best by neoclassical accounts while other aspects are not.

The problems that neoclassical economics pose for either form of IBE can now be seen. For IBUE, there are three objections: (1) the problem of unrealistic assumptions; (2) the application of IBUE rests on questionable empirical assumptions or assumes holistic variables; (3) unifying power is a problematic notion of explanatory power. For (1), the argument is this: Consider an experiential economic situation or system ES. Suppose it to be an actual market or exchange. An NCE model H is proposed as an account of the workings of that system. In order for us to claim that the NCE hypothesis is the IBUE here, Kincaid argues, we need to know (1) that it is indeed a unifying hypothesis, that is, that there are a large range of ES's – other markets as well as a variety of other economic processes – where H serves to explain the experiential behavior we are trying to understand. In addition, we need to know (2) that the hypothesis H is a 'close fit' to the ES, that is, that H is a 'realistic' model. But (2) is an empirical question. So, if and when we decide that H is the best explanation of ES, we are grounding this decision on empirical as well as formal considerations. Now how does this problem differ from our attempt to apply *modus ponens* to justify the inference, on the information provided in the cited example, that 'Tully is mortal'? Presumably, Tully is mortal but the original argument failed to show it because it presumed false information. But, as we saw, this did not

impugn the formality of the principle of inference that we used. Also, when we corrected the premise (an empirical claim) and drew a justified conclusion, no one would say that *this* shows that *modus ponens* is an empirical principle.

As a matter of fact, Kincaid alleges, when we look to the empirical assumptions that economists make, we find them questionable (p. 101). In the terms we are using, the economic models are alleged to not 'fit' the phenomena they are proposed to explain closely enough. I leave this dispute to those with more familiarity with the cases than I. The point is that even if this is true, it does *not show* that IBUE is a suspect, empirically-grounded principle. What it shows, at best, is that the proposed neoclassical economic models do not explain what they are alleged to explain.

Kincaid then looks at three neoclassical models of the firm. Are they explanatory? To be so, they have to satisfy certain empirical assumptions. But, even if they did, to be *neoclassical* accounts, Kincaid argues, they have to be sufficiently similar to 'spot markets – characterized by many buyers and sellers with fully developed preferences and other traits of perfect competition' (p. 104). Notice that this is a very different kind of worry than the one examined above. Here the issue is not the empirical fit between hypothesis and experience but rather a question of whether the so called 'neoclassical' theories of the firm are sufficiently 'close' to some core or essential model of neoclassical economics. In essence, a 'spot market', so defined is a central model (or paradigm) of what a neoclassical model should look like. The problem, as Kincaid sees it, is that the so-called 'neoclassical' theories of the firm may not be sufficiently similar to the spot market model for us to assume that whatever success these theories might have will constitute evidence for the unifying power of neoclassical economics.

What about IBCE? Kincaid sees similar problems here as well (p. 108). The basic idea again is that given two alternative hypotheses, H_1 and H_2, we may opt for H_1 on the grounds that it is the best causal explanation of the data yet H_2 may be preferable on other grounds; it may have greater predictive power, for instance. In such a case, IBCE turns out to be defeasible and hence not a 'foundational' principle (p. 110).

In the particular case of neoclassical economics, two crucial assumptions about 'good explanations' are made (p. 111). First, it is assumed that 'any outcome must be consistent with self-interested behavior'. Second, it is assumed that 'economic institutions and behaviors, so long as they result from a competitive economic process, exist because they are optimal'.

The first assumption comes in 'thick' and 'thin' versions (p. 112). The 'thick' versions, which detail the range of goals and behaviors of agents

are not so plausible and are tied to substantive assumptions about human behavior. The 'thin' versions, which do not spell out the range of goals or behaviors, are more plausible but do not favor neoclassical theories over alternatives.

The second assumption also comes in 'thick' and 'thin' versions (p. 114). The 'thick' versions again are substantive and less plausible as realistic assumptions about human behavior. The 'thin' versions are more plausible but do not favor neoclassical explanations since the details of real markets are often at odds with the characteristics of ideal neoclassical markets (p. 116).

Having lodged these criticisms, Kincaid offers two caveats (pp. 116f.). First, these reservations are reservations about judging theories by appeals to explanatory power. They do not show that neoclassical economics is worse than the alternatives. Second, the reservations should not be read as suggesting that neoclassical economics has no virtues whatsoever. The main point, again, is that we cannot make blanket assumptions about *either* the explanatory power of any given theoretical approach for all applicable situations *or* its commitment to individualism. Each case has to be examined on its own merits.

3. CONCLUSION

There is much more that should prove of interest to both social scientists and philosophers in this remarkable book. The debate between individualists and holists has implications not only for a proper understanding of the social sciences and general scientific methodology but also for issues of morality (the individual versus society and the state), rationality, social contract arguments, and the place of folk psychology in a world of neurobiology. Kincaid's analysis illustrates how complex these questions are. In addition, it provides useful hints about how to formulate a pluralistic hierarchical model of the social sciences that involves genuine multi-level causality and multi-level explanations. All those who are interested in these questions are urged to read this work.

REFERENCES

Becker, G. 1976. *The Economic Approach to Human Behavior*. University of Chicago Press
Becker, G. 1981. *A Treatise on the Family*. Harvard University Press
Danto, A. 1973. Methodological individualism and methodological socialism. In *Modes of Individualism and Collectivism*, J. O'Neill (ed.). Heinemann
Dennett, D. 1969. *Content and Consciousness*. Routledge
Garfinkel, A. 1981. *Forms of Explanation*. Yale University Press
Harman, G. 1965. Inference to the best explanation. *Philosophical Review*, 74:88–95
Howson, C. and P. Urbach. 1993. *Scientific Reasoning: The Bayesian Approach*. Open Court
Kim, J. 1993. *Supervenience and Mind*. Cambridge University Press
Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences*. Cambridge University Press

Kitcher, P. 1989. Explanatory unification and the causal structure of the world. In *Scientific Explanation*. Philip Kitcher and Wesley Salmon (eds.). University of Minneapolis Press
Mellor, D. 1982. The reduction of society. *Philosophy*, 57:51–74
Putnam, H. 1981. Reductionism and the nature of psychology. In *Mind Design*. J. Haugland (ed.). The MIT Press.
Rosenberg, A. 1985. *The Structure of Biological Science*. Cambridge University Press
Rosenberg, A. 1994. *Instrumental Biology, or the Disunity of Science*. University of Chicago Press
Stich, S. 1985. *From Folk Psychology to Cognitive Science*. The MIT Press
Watkins, John. 1973. Methodological individualism: a reply. In *Modes of Individualism and Collectivism*. J. O'Neill (ed.). Heinemann
van Fraassen, B. 1980. *The Scientific Image*. Oxford University Press
Wimsatt, W. 1976. Reductionism, levels of organization, and the mind-body problem. In *Brain and Consciousness*. G. Globus (ed.). Plenum

Michael Bradie

Bowling Green State University

Models and Reality in Economics, Steven Rappaport. Edward Elgar, 1998, vi + 233 pages.

One of the thornier problems in the philosophy of economics is reconciling the explanatory successes of economics (or at very least the perception amongst economists that their theories satisfactorily explain a wide range of phenomena) with the falsity of many of the assumptions that economists make. Rappaport argues that prior attempts to solve this, and other similar problems, have been hindered by commitments to philosophies of science that are misrepresentative of, or inappropriate to, economics. In this book he argues for an approach to economics that gives central place to the construction, elaboration and application of models. And he attempts to flesh it out with a theory of models that explains the many different kinds of uses to which economic models are put. His book is an important contribution to the methodology of economics.

The first half of Rappaport's book is devoted to a careful critique of the various positions in the philosophy of economics that he rejects: McCloskey, Rosenberg, Hausman and Mäki each get a chapter largely devoted to their work. Rappaport's argumentation in this part is rigorous and to my mind largely convincing. Against McCloskey he argues that economists rightly seek knowledge of a theory-independent world. Against Rosenberg, he argues that there is no conclusive evidence that economics has failed to exhibit improvement in its predictions over time and, furthermore, that economics cannot be judged only in terms of its

predictive record. But it is the discussion of Mäki and Hausman that brings the problem of 'false' hypotheses to the fore and so it is worth stating Rappaport's objections to their views.

Hausman and Mäki's positions on the nature of economic theories may respectively be summarized as follows:

(H) Economics is a body of inexact laws that are true *ceteris paribus* of real economies.
(M) Economics is a body of theory true of ideal economic systems arrived at by omission of factors present in real economies.

Neither of these seems to lead to the correct description of all the kinds of general statements found in economic theories. For instance, one of the 'laws' that Hausman identifies, that agents have continuous preferences, is clearly not an empirical claim at all, but an assumption made for mathematical convenience. Something similar is true for the assumption that agents have complete preferences. Completeness of preferences is more plausibly construed as an idealization, but to support this claim we would need to know which are the factors whose omission makes it true in the ideal that preferences are complete. While the failure of some objects to fall to the earth at constant acceleration may be attributed to the presence of air resistance, it seems wrong to say that preferences would be complete save for the presence of some disturbing factor. At very least, economists make no attempt to establish the existence of such a factor or to test for the truth of the completeness claim when influence of the factor in question is negligible.

It is in the second half of the book that Rappaport develops his own conception of economics as a mode of inquiry based on the construction of models. He takes models to be sets of assumptions or axioms and their deductive consequences chosen in order to address particular problems. These include not only the theoretical tasks of explaining and predicting economic phenomena, but also the conceptual and normative tasks of answering questions such as 'Is it possible to ensure cooperative behaviour by appropriate design of social institutions?' or 'What is the optimal rate of taxation?'. Indeed economics is distinguished by the amount of effort that goes into answering questions of the latter kind. With important ramifications, as Rappaport observes, for when a model is constructed to address a conceptual or normative question it is not the way that the world is, but the way that it might be, that is at issue.

Rappaport argues that his modal theory of models gives the best explanation of these many uses of models in economics. At its foundation is a distinction between theoretical and applied models. Theoretical models make claims about hypothetical objects and relations; things that would exist if the conditions postulated by the model were true. Applied

models, on the other hand, are obtained from theoretical ones by interpreting their variables in terms of real world objects and relations. Since theoretical models in effect define their own domain of interpretations (the set of hypothetical objects and relations of which they are true), they are, as he says true 'by legislative postulation'. Consumer theory, for instance, is true of a particular kind of hypothetical object called a consumer – an insatiable utility maximizer – which makes choices amongst another kind of hypothetical object – commodities. Applied models, on other hand, admit of empirical truth, and more significantly, empirical falsity.

The basic distinction between theoretical and applied models seems useful to me: it explains the autonomy of conceptual work and the fact that economic models are often impervious to apparently falsifying evidence (because the evidence is perceived as bearing on the applied model and not the theoretical one from which it is derived). But does recognizing the motivation for the distinction between theoretical and applied models force acceptance of a modal theory? A similar distinction is made by the structuralist theory of models, for instance, which regards (theoretical) models as definitions of complicated predicates which are then applied to particular real-world objects, for example, it takes consumer theory to contain definitions like 'C is a consumer iff C is a utility maximizer and C is insatiable' which are applied to individuals in particular markets making decisions about which goods to buy.

Is there any reason to prefer the modal theory to the structuralist one? Rappaport's argument for the former rests on the claim that it more accurately describes how economists present their models. Now economists do not, I agree, give their models in the form of definitions of predicates. But nor do they present them as being claims, true by legislative postulation, concerning hypothetical objects. Even if economists typically tolerate unrealistic assumptions in their models, they still take them to be representations of the behaviour of real world objects and not hypothetical ones (though in what sense is a tricky question). This may be because economists typically elaborate theoretical models with particular applications in mind. But this is not the point. Both the modal theory and the structuralist one rationally reconstruct, rather than simply describe, the practice of economists.

Which brings us back to the important observation that models are not always, or even typically, put forward as true descriptions of the world. Rappaport explicitly endorses what he calls the partially instrumentalist position that follows from recognizing this. But how is such a partial instrumentalism to be squared with his claim that economists pursue truth? Rappaport's answer is that what they pursue in such cases is second-order truths relating to the usefulness of the theory. But this is just realist sophistry: economists may seek useful

models, and it is axiomatic that, if a model is useful, then it must be true that it is useful, but it does not follow that economists pursue truths of the form' X is a useful model'. Indeed it is misleading to say that they do, in much the same way that it is misleading to say that painters aim at the truth of sentences of the form 'This painting is beautiful'.

In any case the important question, as Rappaport recognizes, is what makes a model useful and what grounds economists' judgements that it is so. It is here that I find his argument to be at its weakest. His first suggestion is that economists will be favourably disposed to models that incorporate assumptions endorsed by prevailing global theories or research programmes. No doubt this is what economists do, but this simply pushes the problem back to justifying the presence of these assumptions in the global theory. Rappaport's suggestion is then that they might be generalizations confirmed by experience or implied by other generalizations that are so confirmed. So his answer amounts to the claim that economists will construct (or find useful) models that contain generalizations or, more frequently perhaps, rough generalizations, that are inductively supported by experience. This makes his position very difficult to distinguish from Hausman's. For what else are these inductively supported rough generalizations other than the inexact laws that Hausman takes to be characteristic of economics?

More to the point, the only justification we are being offered for the usefulness of a model is the truth or rough truth of the generalizations it employs. But this is not a complete answer to our problem because it is clearly not sufficient for a model to be useful that it contains some true or roughly true generalizations. In any case, models do not just consist of such generalizations, but also typically contain assumptions specific to the domain being modelled, as well as various simplifying assumptions. We need some account of what justifies their presence. Indeed the problem with which we began concerns precisely the presence of assumptions that are not just false, but in many cases not even presented as being true or approximately true. But despite recognizing the importance of the question for an understanding of conceptual and normative models, Rappaport simply fails to address it.

This is not to say that the question cannot be answered in a manner consistent with Rappaport's framework. Quite the contrary. A plausible standard for the kind of assumptions invoked in models exploring conceptual questions is that it be practically possible to make them true or approximately true. For that is what is required if the model is to have policy implications. Indeed, it seems to me economists are very sensitive to considerations of that kind. They not only recognize that models may be better or worse approximations of the real world, but that they may describe circumstances that are more or less difficult to bring about (or to avoid) by the intervention of agents of one kind or another. (There is a

similarly motivated interest in the stability of such circumstances, once obtained).

I have argued that Rappaport does not in the final analysis give sufficient grounds for preferring his modal theory of models to the structuralist theory. More importantly he does not give a full account of what makes a model a good (or better) one. But this is an important omission in a book that has recognizes the multiplicity of uses of models. This is not to deny the value of Rappaport's book. It presents a well argued and coherent position within the philosophy of economics, which by placing models at the centrepiece of analysis does more justice to economics than many other prevailing accounts. I highly recommend this book to anyone interested in the philosophy of economics.

Richard Bradley

London School of Economics

Incommensurability, Incomparability, and Practical Reason, Ruth Chang (ed.), Harvard University Press, 1998, 303 pages.

One focus of dissatisfaction with standard models of decision making has been the assumption that a rational agent must have a complete linear ranking of the outcomes open to her. Such agents seem very different from human beings, or, to put the point more carefully, they seem to represent a state that humans can only approximate to as the result of a good deal of work on their preferences, a process that intuitively seems to be susceptible to rational criticism. You can do it more or less well. So, at the very least, standard models leave out something important and interesting. Current interest in philosophy in these questions is largely a product of dissatisfaction with the 'Humean' dogma that there can be no rational deliberation about ends. In economics, the interest largely stems from a general suspicion that too heavy idealization in microeconomics may close off interesting ideas about the behavior of consumers and investors. The philosophers' and economists' paths crossed at a chateau in Normandy, for an interdisciplinary conference. This book consists of descendants of some of the conference papers, largely those by philosophers, many of them taking account of points made by others of the collection, plus some commissioned pieces. Ruth Chang has added a very helpful and clarifying introduction.

https://doi.org/10.1017/S0266267100240182 Published online by Cambridge University Press

Each paper makes a point. There is a fair amount of overlap between the papers. The points argued for fall into four rough categories.

The nature of incomparability. There is considerable divergence among the authors about the nature of the phenomenon. The simplest interpretation of incomparability is to take it as a failure of trichotomy, the principle that A is preferred to B or B to A, or they are indifferent in preference. Even then there are choices to make. Are we dealing with the merely partial ordering of specific outcomes as ranked by a particular agent, or with some feature of preferences between general desirable and undesirable features of the world, such as liberty and well-being? But trichotomy, indeed the structure of preference orderings, may not be the issue at all. Something more subtle about rationality or the process of thinking through a choice may be crucial. To some extent these differences are just claims for ownership of the word 'incomparable'. They also involve competing claims about which phenomena are more basic and most worth trying to understand. David Wiggins seems to be working towards a conception of incomparability in which the focus is on the difficulty and seriousness of making comparisons, and their elusiveness, rather than on failure of trichotomy. It would make sense to call this 'incommensurability' as there is an intelligible metaphor with Pythagorean commensurability of rational and irrational numbers within the complete ordering of the real line. Elizabeth Anderson uses a simple characterization of incomparability as failure-of-trichotomy and argues that to do justice to it we need a radically different model of practical reason. Ruth Chang points out the variety of possible preference structures and begins the task of relating these to characteristic difficulties of comparison and decision.

Does it exist? To some it is just obvious that the standard idealizations do violence to the shape of our preferences. Thus, David Wiggins, Elizabeth Anderson, Joseph Raz, Elijah Millgram, Charles Taylor, Michael Stocker, and Cass Sunnstein simply work from the assumption that their construal of incomparability represents something real. Donald Regan plays the role of, as he puts it, 'the ''designated eccentric'', appointed to take a position no one else would touch with a barge pole'. He defends – with clarity, sense, and good humor – the suggestion that there really is no such thing as incomparability of desires, values, or outcomes. His arguments come down, in the end, to the observation that often when we are unable to compare two values or outcomes more reflection will produce the missing comparison. He points out that it is often morally required that we try to do the thinking that may result in a comparison. While this is undeniably true, it misses two basic points. First, sometimes when we think about two at first sight comparable outcomes or values

we become *less* certain that we can rank either of them above the other, while remaining torn between them. Second, as Chang, Anderson, and Sunnstein point out, sometimes we are morally obliged to try to avoid thinking through a problem by finding a trade-off between the competing factors.

John Broome argues for a position that at first sight seems much like Regan's. The effects of real incomparability between the values of outcomes may be the result of ignorance-induced vagueness. In fact, though, what Broome argues, convincingly, is that even if the values of outcomes – their objective degrees of goodness – are thoroughly comparable, decision-makers will still have to deal with the fact that they are often unable to determine these degrees of goodness precisely enough to treat those outcomes as totally ordered. Regan could use this conclusion to explain the appearance of incomparability. But his opponents could also use it, to explain the need for taking incomparability as a serious issue for decision-makers, whatever the ultimate and perhaps unknowable structure of good.

Non-maximizing patterns of reasoning. If incomparability is inescapable then we need decision-making procedures which do it justice. Elizabeth Anderson suggests that what we need is an 'expressive theory of rational choice' in which the fact and manner of choice is constitutive of the ends that choice aims at and balances. The general flavor seems to be anti-maximizing, though with a much more pluralistic quality than many deontological accounts. As she points out, Kant's ethics are expressivist in that the only ultimate value is reason itself, but one can make the rational manner of a decision part of the value of an outcome without either seeing reason as homogeneous or making it into the sole desideratum. She does not consider cases analogous to the 'murder to prevent more murder' cases that separate hard core anti-maximizers from the rest. She argues nicely for an extreme variety of patterns of value and of valuation.

Joseph Raz argues that the required reasoning does not operate with desires at all. He points out that when someone approaches a hard decision they ask themselves what they should prefer, rather than what their inclinations or whims are. As a result, many of our wants are based on our conclusions about what is valuable, rather than the other way round. A complex structure of values is, for Raz, at the heart of practical reasoning, and must be grasped both when making non-trivial decisions and when understanding the choices of others. Raz's arguments and examples do not actually make much essential use of incomparability. Much of what he says would hold true of other circumstances in which choice is difficult. One of Raz's conclusions is endorsed by several other authors in this volume: in making sense of non-trivial decisions we need

something richer than 'desire' and 'preference'. Raz distinguishes between goals, desires, reasons, and urges. Regan and Stocker make similar points explicitly, and others hint at them.

Elijah Millgram discusses the process of making preferences comparable. Or more generally, since this is not an argument for general comparability, the process of fitting one outcome into a preference-structure alongside others, or of reorganizing an existing structure. According to Millgram some aspects of this process will inevitably occur during the act of decision rather than in preceding reflection. Millgram's main aim is to dissuade us from making such comparisons in our heads, by a priori or imaginative means. It is essential, he thinks, to form our preferences by reflection on our experiences. And when we see how we can do this we become more optimistic about the possibilities for rational preference dynamics.

Links with moral philosophy. When we stop thinking that rational decision aims to maximize the satisfaction of whatever desires the agent happens to have, we begin to erode the distinction between decision theory and moral philosophy. The best choice is the one that gives you most of what it would make most sense to want. Or what it would be rational to want, or even what you should want. It is not surprising then that several authors argue that reasoning with incomparables, even when it does not involve balancing of one person's interest against another's, has characteristics sometimes attributed to morality. An aversion to maximization is found in several papers, sometimes not clearly distinguished from an aversion to the arbitrariness of desires. Charles Taylor argues that the resources we bring to problems of incomparability are those of thinking through the larger structures of our lives. He says that we have resources that are not acknowledged in philosophy for making sense of our lives, and which are essential when we are faced with deep incomparabilities. He does not say very helpfully what these resources are. Stephen Lukes, Ruth Chang, and Elizabeth Anderson, stress the moral importance of not making easy comparisons of the value of, for example, friendship and money. And James Griffin and Cass Sunnstein stress the social and legal dimension of the point: we can acknowledge publicly that something has a certain kind of value by building obstacles to simple trade-offs between it and other things into our shared practices.

There are remarkably few outright disagreements among these papers. Even Regan accepts that most people most of the time cannot rank many alternatives open to them, and may have to make decisions before they can find or impose a suitable better-and-worse ordering. I believe that this appearance of harmony is in part based on a mistake. The mistake is to think that if you accept the existence of incomparabil-

ities in preference orderings you are driven away from maximization. This is simply not so. Outcomes can be assigned values over a partially ordered set, and the formal machinery of utility maximization can be formulated more or less unchanged. Of course, very often incomparability in outcomes will then generate incomparability in acts. There is nothing awful about this, though it does invite us to formulate supplementary principles to say what an agent should do when two available options are incomparable and not dominated by any others. (The simplest principle is: do either.) Once we realize this, that incomparability is not by itself a weapon for attacking Humeans and consequentialists, it becomes easier to separate out the issues that do have a bearing on questions about maximization. They are – I claim, with the support of several papers in this book – issues about the formation and evolution of preferences, and issues about limited rationality. In particular they concern decision making when the decision is needed soon but thinking out a solid ordering of the outcomes will take longer. The crucial fact, I think, is that faced with the task of making sense of our preferences all human rationality appears very limited. Discovering what to want is a very hard job, and takes as much time and intellect as we have to give to it, so that most decisions have to be made on the basis of a very inadequately thought out set of preferences. Incomparable desires invite us to enter a difficult long-term process of preference revision. (They are not the only invitation, of course.) They thus reveal the need for two distinct kinds of principle. One kind concerns the evolution of preferences: the long-term thinking out of what we should want. The other kind concerns decision making with inadequate materials: the materials are always inadequate. Most of the papers in this book contain suggestions about principles of both kinds, usually without separating them very clearly. The suggestions are not easy to turn into definite principles, or even into useful guidance for people faced with hard choices. There is a lot of work to do.

Adam Morton

University of Bristol

Just Playing: Game Theory and the Social Contract Vol. 2, Ken Binmore. MIT Press, 1998, xxiii + 589 pages.

Just Playing is the second volume of Ken Binmore's *Game Theory and the Social Contract*. The first volume was entitled *Playing Fair*. These titles

https://doi.org/10.1017/S0266267100240182 Published online by Cambridge University Press

refer to the two poles of the theory, the Game of Morals and the Game of Life. Binmore models the Game of Life as an indefinitely repeated game. Players of the Game of Life just try to optimize payoffs – they are 'just playing'. Players of the Game of Morals interpose between the stages of the Game of Life, stages in which any player may demand renegotiation behind a veil of ignorance. The negotiations behind this veil determine how subsequent stages of the game of life are to be played unless further renegotiation is invoked. Players of the Game of Morals are 'playing fair'.

The resolution of the tension between the Game of Morals and the Game of Life – between 'playing fair' and 'just playing' is the central concern of Binmore's theory of the social contract. While other authors (Harsanyi, Rawls) have held that analysis of rational decision behind the veil of ignorance yields prescriptive answers to moral questions different from the question of rational action in the game of life, or that the moral question can be answered without the veil of ignorance (Gauthier, Scanlon), Binmore wishes to show how the veil of ignorance can be invoked by those who are 'just playing' so that they end up 'playing fair'. Binmore wants to show how the Game of Morals may ultimately be imbedded in the Game of Life; how rational players may reach an equilibrium in the game of life that is also an equilibrium in the Game of Morals.

This is possible only if playing the Game of Morals is of mutual benefit to the players. In Binmore's model the Game of Life is a repeated game. Then 'folk theorems' for repeated games (discussed extensively in Chapter 3 of this book) show how outcomes attainable by cooperation in the stage game can be approximated by subgame perfect equilibria in the repeated game. If our society is not Pareto-effecient, we can mutually benefit by renegotiating a social contract and moving to a new equilibrium. Incremental Pareto improvements can be pieced together to form a path leading to a Pareto-efficient state. Binmore sees the use of the Game of Morals as a coordination device that has evolved as a way of agreeing on such Pareto improvements. If so, it can be in everyone's best interest to participate in the Game of Morals. Players simultaneously seek an equilibrium in the game of morals and an equilibrium in the game of life.

Already it is evident that there is something new and interesting in this book for moral philosophers to consider. There is more. Rational decision behind the veil of ignorance requires you to, in effect, *forget who you are*. You have to think as if you had an equal chance of coming out as person of type A or person of type B. You must, then, be equipped with *empathetic preferences* for persons of type A and persons of type B. You must be able to make judgements of the type THIS, if A is better than THAT, if B. Such preferences have been previously discussed by Suppes,

Arrow, Sen and Harsanyi, as 'extended sympathy' preferences. Binmore adds something new to these discussions. He has a theory of the co-evolution of empathetic preferences and the device of the veil of ignorance, discussed in Chapter 2:

> I think that we have empathetic preferences at all only because we think of them as inputs when using rough-and-ready versions of the device of the original position to make fairness judgements in real life. Insofar as people from similar cultural backgrounds have similar empathetic preferences, it is because the use of the original position in this way creates evolutionary pressures that tend to favor some empathetic preferences at the expense of others. In the medium run, the result is that everybody in a society tends to have the same set of empathetic preferences. (p. 178)

If empathetic preferences agree they form a standard for a kind of interpersonal comparison of utility. Binmore's theory contrasts with Harsanyi's account, which appeals to psychology for interpersonal comparisons, and to Rawls, who uses primary goods to get the interpersonal comparisons that are required by his theory. That is not to say that these different kinds of accounts are incompatible with one another. Primary goods may loom large in psychology, which was shaped by evolutionary processes.

There is another feature of Binmore's approach that gives it its special flavor. That is, Binmore's refusal to assume any commitment on the part of his actors. Thus, after they may reach an agreement behind the veil of ignorance, neither is obliged to carry it out. Self-interest rules at each stage of the process. Once outside the veil, if a player does not like the agreement, she can immediately demand renegotiation. This puts a stringent control on what sort of agreement it makes sense to negotiate behind the veil. If the agreement is to be carried out, the expected utility of renegotiation cannot be higher for any player than the expected utility of carrying out the agreement.

Suppose the problem is just how to distribute a windfall of $100 between A and B. We agree that here utility just equals money. You and I, behind the veil, act as if we have probability 1/2 of being A or being B. When we come out from behind the veil mother nature will flip a coin and decide. Now, if we only maximize expected utility behind the veil, any distribution between A and B will be equally good. But we know that any distribution other than equal shares will not fly. If B gets $40, and A gets $60, then as soon as we emerge from behind the veil and find out who is A and who is B, the one of us who is B will demand renegotiation. Renegotiation behind the veil has an expected payoff of $50 for each of us and $50 is better than $40. This is how Binmore pulls the 'maximin rabbit' out of the 'Bayesian hat' in Chapter 4 (p. 437).

However, Binmore notes that if the payers were able to commit

behind the veil of ignorance, or if there were some external enforcement mechanism that would hold them to whatever contract they arrived at behind the veil, then the theory would give utilitarianism – as Harsanyi argued – rather than egalitarianism. The crucial difference between Binmore and Harsanyi depends on the ever-present possibility of renegotiation of the social contract, and the impossibility of commitment.

But although Binmore pulls a kind of egalitarian rabbit out of the hat, it is not quite Rawls's rabbit. That is because Binmore's egalitarianism is tempered by his incrementalism. For Binmore, what is up for negotiation is not what you already have, but rather how to share the fruits of some mutual improvement. It is the incremental utility that can be generated by a new contract that you are bargaining over. If a poor man and a rich man go for a walk and they find $101 in the street, they may negotiate about the $101 behind the veil of ignorance, but not about total wealth.

Incremental egalitarianism is egalitarianism with respect to utility, not money, and this may lead to curious consequences. Suppose that the rich man and the poor man agree that their rather different lives are of equal value, but less than perfect. They need different things. The rich man lacks the fulfillment that comes from sustained philosophical reflection. The poor man could use a little cash. They both agree that the utility that the poor man would get from an additional dollar (within the range of possibilities under consideration) is 100 times the utility that the rich man would get from an additional dollar. Utilitarians would view this as an argument that the poor man should get it all – and the classical utilitarians used this sort of consideration as an argument for social reform. But it is hard to see how incremental egalitarianism can escape the conclusion that the rich man gets one hundred dollars and the poor man gets one – giving them equal increments in utility.

Binmore would say that in the 'medium run' the poor and rich man would not share the utility judgements that I have postulated. He argues that cultural evolution would have reshaped empathetic preferences so that bargaining behind the veil of ignorance would agree with the Nash bargaining solution applied outside the veil. (This is one place in the book where I would have liked to see an explicit dynamical model of how cultural evolution is supposed to operate.) Morality, in this framework, only has bite in the 'short run' and accordingly, the rich man-poor man story should be taken as a short-run example.

The stark contrast between utilitarianism and egalitarianism (renegotiation and commitment, Rawls and Harsanyi) is blurred if there is a more-or-less effective community enforcement of fairness:

> When fairness norms are employed in practice to settle day-to-day coordination problems, it is seldom the case that no source of external

https://doi.org/10.1017/S0266267100240182 Published online by Cambridge University Press

> enforcement is available. Imagine, for example, the weight of public disapproval that would follow if a captain were to refuse to honor the fall of the coin used to determine the initial choice of ends at the soccer match, and were to insist instead that it be tossed again until it fell in his favor! (p. 432).

Binmore insists, however, that we cannot assume community enforcement has always existed. It, itself, had to evolve over long periods of time.

Does one have to know game theory in order to read this *Just Playing*? No, the author develops the requisite theory along the way. The introduction introduces the reader to Nash equilibrium and subgame perfection. Chapter 1 is an extended tutorial on bargaining theory. Chapter 3 discusses the folk theorem for indefinitely repeated games and renegotiation-proof equilibria. An innocent but diligent reader will learn some game theory along the way. One who works through both volumes of *Game Theory and the Social Contract* with a game theory text (such as Binmore's *Fun and Games* or Myerson's *Game Theory: Analysis of Conflict*) handy will learn more.

It is written in Binmore's lively, irreverent style. Plato and Kant, as well as various contemporaries, are skewered and roasted. The organization is holographic. Each chapter refers forward and back, until – after experiencing all the shifting views – the reader is left with a conception of the theory in the round. It is worth the effort.

Game Theory and the Social Contract (*Just Playing*, together with its companion volume, *Playing Fair)* is a major contribution to social philosophy. Every serious student should study it. Every serious theorist will have to come to terms with it.

Brian Skyrms

University of California, Irvine

Preferring Justice: Rationality, Self-Transformation, and the Sense of Justice, Eric M. Cave. Westview Press, 1998, xiv + 183 pages.

Following Hume, but simplifying somewhat, let us say that a group of individuals is in *the circumstances of justice* when their situation is such as to require a significant degree of self-restraint, and interpersonal cooperation, if they are to be better rather than worse off, over time, so far as advancing their own interests is concerned. We might then say, following Eric Cave and any number of other recent writers, that coordinating the relevant cooperative agreements in these circumstances, and identifying who needs to restrain himself in what ways, and when,

creates a need for what one might call *a conception of justice*: a rule or set of rules by means of which the relevant individuals resolve the coordination problems the circumstances of justice create.

Suppose one believes, as Hume did, and as Cave appears to do, that in one's own society there is general agreement on a particular conception of justice, and that this conception is sufficiently rich, and sufficiently well-articulated, as to make it fairly clear who is required to do what, and when, by way of self-restraint, if the goods that social cooperation makes possible are to be attained. One problem one might then want to confront – Hume states this problem eloquently in his second *Enquiry*, as did Hobbes before him in Chapter XV of *Leviathan* – is the problem of explaining why a rational individual, thinking only of his own well-being, would go along with the dictates of justice in situations in which the best available evidence suggests that doing so will be inconsistent with the maximization of his own best interests.

Cave's aim in the book under review is to solve this problem. Central to his proposed solution is the notion of *the sense of justice*. To have the sense of justice, on Cave's view, is to be disposed to comply with the rule or rules that articulate the requirements of justice in one's society, and to be disposed to do this not for instrumental reasons but because one sees doing so as intrinsically valuable. Individuals with a genuine sense of justice would have good reason to act in accordance with their society's conception of justice, on this view, and would, when thinking clearly, see that they have good reason to do so, even when, on a certain way of thinking, it appears that to act justly will be inconsistent with the maximization of their own interests. This would be so because to have the sense of justice, as Cave understands it, is to prefer to act justly, rather than unjustly, and to prefer so to act because of what one sees as the intrinsic desirability of choosing justice over injustice, regardless of whatever other goods one might acquire by acting against the requirements of justice.

Why would rational, basically self-interested individuals have such an intense preference for acting in accordance with the requirements of justice – a preference sufficiently strong, if Cave is right, as to make it nearly always irrational to act otherwise (i.e., against the requirements of justice), regardless of the beneficial consequences to oneself of doing so? The answer, according to Cave, is simple enough, though showing that it is the (correct) answer admittedly takes a bit of doing.

Instrumentally rational individuals in circumstances of the relevant sort, Cave argues, will see, to begin with, that they will be better off if most of them have and regularly act on the requisite sense of justice – if most of them have, that is, a disposition to comply with the rule or rules that constitute their shared conception of justice and to do so because they see doing so as intrinsically valuable. But, seeing this, Cave

continues, such individuals, though initially lacking a sense of justice, will take steps to see to it that they come to have a sense of justice of the requisite sort. Specifically, they will see the desirability of bringing it about that they prefer acting justly over not acting justly, and that they prefer this as an intrinsic end, and so they will do their best to bring it about that they have, or that most of them have, the relevant preferences.

Cave is aware of the fact that he must achieve at least two very different tasks if he is to make good the argument just sketched: (i) he must convince us that instrumentally rational individuals of the sort he is imagining will see the collective and individual desirability of bringing it about that they have the preferences the sense of justice requires them to have on his view; and (ii) he must convince us that such individuals will in fact be able to bring it about that they have these preferences. Curiously, though, he tries to achieve, at least in a serious way, only the first of these two tasks in his book, leaving one with nothing but questions about the tenability of anyone's ever achieving the second task. (He devotes one-hundred-and-fifty-six of the book's one-hundred-and-seventy-three pages of text to the first task and fewer than eight to the second.)

By way of discharging the first task, Cave argues, first, that individuals of the relevant sort would see the collective rationality of bringing it about that they have (or that most of them have) a sense of justice of the sort his overall argument requires, and he argues, secondly, that these same individuals would see the individual rationality (because they would see the desirability) of bringing it about that their individual preferences are such as to strongly dispose them to act in accordance with the requirements of justice, even in cases where, these strong preferences to one side, it would be rational to act against the requirements of justice.

Although the general structure of Cave's arguments in this connection will not be surprising to those familiar with the very extensive recent literature on these matters, the details are sufficiently novel, in my view, and sufficiently interesting, as to make a look at them eminently worthwhile. Cave does not, it is true, say very much about the many alternative treatments of these same issues that can be found in the recent literature: in fact, he spends just sixteen pages 'disposing' of the competing views of Sen, Gauthier, McClennen, MacIntosh, and others, and he devotes not a single word to the views of historical Greats like Plato, Hobbes and Hume. Nonetheless, he takes quite seriously the need to achieve the first task mentioned above, and he shows a certain flair, both technically and philosophically, in attempting to achieve it (in attempting to show, that is, that rational, non-tuistic individuals situated in the relevant circumstances would do their best to bring it about that they have the preferences Cave's overall argument requires).

Unfortunately, Cave spends, as I have already indicated, almost no time at all attempting to achieve the second task his overall argument requires him to achieve – namely, showing that rational, basically self-interested individuals would be *able* to change their intrinsic preferences in the way his argument for 'preferring justice' requires. In a chapter that, with a page-and-a-half of footnotes, is just eight pages long, he essentially just asserts that individuals of the relevant sort would be able to do what his argument requires them to do and that they would be able to do it because of the efficacy of a phenomenon he calls 'habituation' (very briefly, he claims that by regularly acting in accordance with justice, even when one can in one clear sense do better for oneself by acting against the requirements of justice, one will come, given certain other preferences, to have an intrinsic preference for acting justly rather than unjustly, despite the advantages of going the other way). There is no discussion at all of the available empirical literature on preference-change – in fact, there are no references to this literature, even in his notes – and there is just the barest beginning of an attempt to discuss the many difficulties one might raise, empirical studies to one side, with the suggestion that one might, with practice, fairly easily change one's intrinsic preferences in the way Cave's argument requires.

This, of course, will not do. Indeed, the situation is even worse than I have suggested, inasmuch as Cave seems at times to be deeply confused about what it is he has to prove in this connection: while at times, in this short chapter, he appears to recognize that what he has to show is that (and how) individuals of the relevant sort could in fact change their intrinsic preferences in the requisite ways, at other times he gives clear evidence of thinking that all he needs to show is that it would be logically possible for individuals of the relevant sort to change their intrinsic preferences in the requisite ways (see especially the second full paragraph on page 163). This is really too bad, and, the howler just noted to one side, it suggests that this is really just half the book Cave wants and knows he needs to write, rushed into print prematurely for reasons one can only guess at.

Nonetheless, I think this is a book those interested in the relevant issue(s) will want to examine. It makes, as I have indicated, a serious attempt to achieve the first of the two tasks it needs to achieve, and in the process suggests arguments that are both novel and challenging. Perhaps its failure to address seriously the other task it needed to achieve will stimulate others to try to do better.

Daniel M. Farrell

The Ohio State University

https://doi.org/10.1017/S0266267100240182 Published online by Cambridge University Press