Erasmus Journal for Philosophy and Economics,
Volume 3, Issue 2,
Autumn 2010, pp. 111-118.
http://ejpe.org/pdf/3-2-br-3.pdf

Dancing at gunpoint. A review of Herbert Gintis's *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton: Princeton University Press, 2009, 304 pp.

TILL GRÜNE-YANOFF
University of Helsinki

The bounds of reason seeks to accomplish many things. It introduces epistemic game theory, discusses other-regarding preferences in games, offers an evolutionary model of property rights, and proposes a plan to unify the behavioural sciences. Most notably, it is a plea for the importance of human nature and sociality for the determination of strategic behaviour on the one hand, and a defence of traditional decision theory on the other.

Being *normatively predisposed* by their nature, human players accept social norms as correlation devices that *choreograph* a correlated equilibrium. While social norms put on the dance, epistemic game theory is driven by the "cannons of rationality" (p. 83), as Gintis puts it in one of the many and sometimes hilarious misprints. Traditional decision theory is "mostly correct" (p. 246), and Gintis relies largely on its support for solving games. Thus the choreographer is restricted to where the cannons cannot reach. Game players are dancing at gunpoint here—with important consequences for the proposed unification of the behavioural sciences.

But I am jumping ahead. The main part of *The bounds of reason* concerns a decision-theoretic approach to game theory. Its purpose is to investigate the (Bayesian) epistemic basis for central solution concepts, both as a justification of what is reasonable, as well as a derivational basis for predicting what is actually observed. This Bayesian rationality forms Gintis's "cannon of rationality", which he aims at various game theoretic solution concepts.

The first victim of this artillery is the assumption of common knowledge of rationality (CKR). Gintis argues that CKR is neither derivable from Bayesian rationality nor can it be epistemically justified on its own. It therefore cannot function as a premise of game theory, but must rather be interpreted as an "event" that may or may not occur

(p. 100). This argument wreaks two kinds of collateral damage. First, rationalizability in normal form games loses its epistemic justification. Without CKR, players do not necessarily eliminate all unrationalizable strategies. This is indeed a relevant possibility, as Gintis illustrates with a number of intuitive and experimentally supported cases. Second, subgame perfection is undermined. With CKR demolished, no alternative epistemic justifications of subgame perfection are available (p. 120), "it is reasonable to assume Bayesian rationality, avoid backward induction, and use decision theory to determine player behavior" (p. 112).

Gintis's next target is the Nash equilibrium (NE) itself. The sufficient epistemic conditions for NE in games with more than two players are common priors and common knowledge of conjectures. Gintis employs two different ordnances to destabilize these foundations. First, using modal logic, he argues against the claim that any event self-evident to all members of a group is common knowledge. This is true only, he shows, if one assumes that the way each individual partitions the universe is known to all (pp. 152-153), but that is a much stronger assumption and more likely not satisfied. Second, he argues that the sufficient conditions for NE are a kind of agreement theorem à la Aumann (1976). Agreement theorems of this sort have implausible implications, for example that rational, risk-neutral agents with common priors and common knowledge of posterior probabilities would not trade assets. Thus, Gintis concludes, common knowledge (or common priors or both) are widely violated, putting the applicability of the Nash equilibrium in doubt.

Having thus established the field of fire, the dancing can begin. The tune is set by Aumann's *correlated equilibrium*, which Gintis considers "a more natural solution concept than the Nash equilibrium" (p. 44). The idea is that an existing game is expanded so that "Nature" first gives a publicly observable signal. Players' strategies assign an action to every possible observation. If no player has an incentive to deviate from the recommended strategy—assuming the others do not deviate—the distribution is called a correlated equilibrium. Nature, then, is the choreographer who guides players' choices in areas where the cannons cannot reach.

Correlated equilibrium only requires rationality and common priors, while Nash equilibrium requires stronger premises. But where do the common priors come from? Drawing on Vanderschraaf's (1998) analysis, Gintis defines a *symmetric* reasoner as an agent who can infer, from his

own conclusions about a state of affairs, the conclusions of other players. In a group of Bayesian rational symmetric reasoners, mutual knowledge of an antecedent implies common knowledge of the conclusion (Theorem 7.2). Presumably (although this remains vague) the possibility of a symmetric reasoner depends on social properties like cultural norms. Thus, it is social properties that make common priors possible.

However, if the correlated strategies involve multiple strategies with equal payoffs, then players have no incentive to follow the choreographer's instructions. This is where social norms come in more explicitly: if norm-conforming behaviour is a correlated equilibrium, then players will choose the corresponding correlated strategies (Theorem 7.3). Gintis bases the main message of the book on these two results, namely that the decision-theoretic approach to game theory is incomplete: it requires more than "mere player rationality" to solve (at least some) games. Where the cannons cannot reach, ballet is supposed to lead the way.

Gintis stresses the various methodological implications of this result. First, he argues, this implies the rejection of methodological individualism: agent behaviour depends on social emergent properties—common priors and common knowledge—that cannot be analytically derived from a model of interacting "merely rational" agents. Second, reason is "socially bound" by the existence of social norms that cannot be explained by individual rationality itself.

With game theory thus circumscribed, Gintis proposes a unification plan for the behavioural sciences. He identifies four incompatible models: the psychological, the sociological, the biological and the economic, "all four […] flawed" (p. 221). Out of this flawed mess, Gintis sets out to forge a correct whole. Maybe not surprisingly, decision theory forms the core of this unified approach. Gene-culture evolution and socio-psychology are to detail the shape of the utility function, sociology is also supposed to explain the existence and form of the normative choreographer, and complexity theory deals with emergent properties.

I find this proposal for unification not very convincing. Gintis shoots wide, leaving out so many details that it is difficult to see what a unified behavioural science would look like. How, for example, is complexity theory to deal with emergent properties? The author does not say, but apparently could not resist throwing in buzzwords like these, either.

Further, if the four models are incompatible, how does Gintis seek to make them compatible? He only says that in the unified discipline, sociological and economic "forces" will complement each other (p. 242). But such a divided-domain perspective is not so new. Mill (1844 [1836]) long ago characterised economics as investigating certain causal tendencies; the result of these partial investigations had to be synthesized with investigations from other disciplines in order to explain or predict real world phenomena. The question remains how these forces are to be properly delineated.

But maybe one should instead interpret Gintis as proposing a division of labour. For example, as Gintis suggests, gene-culture evolution and socio-psychology are to detail the shape of the utility function, which decision theory then bases its work on. But that again is not news—economists have lived with this so-called Robbins-Parsons division of labour for most of the latter half of the twentieth century (Hodgson 2008). My impression is that this division has become increasingly obsolete, both because economists themselves have become more interested in the form of the utility function, and because some psychologists and sociologists have moved away from seeing utility functions as central to behavioural explanations. Indeed, most of the empirical research into the form of the utility function over the last 20 years or so has happened *within* economics, not in sociology or psychology. Furthermore, there is considerable controversy about this research. Some economists insist on expanding the utility function. Gintis, for example, believes that "internalised norms are arguments in the preference function that the individual maximizes" (p. 233), and he thinks that this research will yield "a pattern of human attributes that can likely be subjected to axiomatic formulation much as we have done with the Savage axioms" (p. 144). Yet even some of Gintis's behavioural colleagues are cautious about attributing such "individual propensities" (Loewenstein 1999, F31), and instead suggest associating behavioural traits with certain contexts. Others have argued that the attribution of fairness preferences and similar is not borne out by the empirical data, and that sensitivity to norms cannot be explained by including new terms in the utility function (Bicchieri 2006). Thus, Gintis's vision of unification is likely to meet resistance even in his home science, economics.

This holds *a fortiori* for psychology and sociology. Many evolutionary psychologists, for example, prefer explaining behaviour as

the result of context-dependent, adapted heuristics, rather than as the outcome of the optimization of a utility function under constraints (Gigerenzer and Brighton 2009). Gintis brushes these differences aside as mere preferences for procedural over as-if models (p. 236), but the difference cuts deeper than a mere question of realisticness. First, focusing exclusively on the model that best fits the data increases the danger of 'overfitting'. The more flexible a model, the more likely it is not only to capture the underlying pattern in the data, but also unsystematic patterns such as noise. Thus, it may be methodologically prudent to restrict oneself to parsimonious procedural models rather than to as-if models with a large number of free parameters. Second, when deriving normative conclusions from decision models, the way deliberation procedures are represented often matters. Gintis himself stresses this for the case of epistemic game theory, where a definition is deficient because it "does not tell us how to find the set that satisfies it" (p. 91, see also p. 195); but he apparently applies different standards for the underlying decision theory. Thus, many psychologists and sociologists may not be willing to accommodate themselves to Gintis's unification proposal: they may think that Gintis overstates the reach of his cannons, and refuse to limit their dance to those few areas where Gintis does not claim firing rights.

This brings me to Gintis's view of the status of decision theory itself. When asserting the correctness of decision theory, it is not clear whether Gintis means this in a normative or a descriptive sense. When discussing decision and game theory in the book, he refers both to "plausibility" (p. 90) and "common sense" (p. 109), as well as evidence from behavioural experiments. At least in a descriptive sense, the correctness claim is controversial. Over the last few years, mainstream economists have redoubled their efforts to rationalize choice that violates the weak axiom of revealed preference (WARP) (e.g., Bernheim and Rangel 2009; and references therein). I cannot see how genuine WARP violations can be compatible with classical decision theory, unless one gives up on the idea of revealed preferences altogether.

Gintis may be inclined to do so, since he suggests that preference inconsistencies can be resolved by using a "more complicated choice space" (p. 9)—i.e., including more parameters in the utility function. This sits well with Gintis's professed as-if perspective of mental models: optimization models are only employed to describe behaviour, not to make claims about the actual psychological set-up of agents (p. 236).

But it chafes uncomfortably against the claim that economic models are supposed to be "testable" (p. 129), and that game theory follows the "hypothetico-deductive method" (p. 223). Without tight constraints on how the choice space can be re-described, how can one test these models? The danger is that re-description continues until all existing data is fitted, and then the form of the utility function has become so weighed down with parameters that no meaningful tests are possible anymore.

Gintis further offers an evolutionary defence of decision theory, yet the few references he gives model the evolutionary context under highly specific conditions. The danger here is always that such models present just-so stories without sufficient robustness. But just take Gintis's own case for transitivity: an organism with an optimized brain, he argues, chooses transitively. That is, if that organism, choosing between pairs of alternatives, chooses A over B, and B over C, then it will also choose A over C when both are available (p. 235). But at the same time, Gintis stresses the dependence of preferences on contexts and current states. This should then also hold for evolution: when choosing between A and B, or B and C, the selective pressures may be different than when choosing between A and C. Hence natural selection does not necessarily entail preference transitivity.

With the fire power of decision theory seemingly somewhat less than Gintis claims, the rationale of epistemic game theory may shift. Gintis presents the reader with a strong contrast between well-founded Bayesian rationality and the baseless assumptions of classical game theory. Recall his conclusion that CKR is "an event, not a premise". Presumably, this means that CKR is sometimes false of real-world situations, and because premises must be true in general, CKR cannot function as a premise, and should be discarded from the game-theoretic toolbox. But if decision theory itself is not as well-founded as claimed, then presumably assumptions like preference transitivity cannot function as premises, either. But that would be an absurd conclusion. Rather, it seems that Gintis is operating with a very narrow view of modelling methodology, in which model assumptions must be true to be acceptable. Given that the contrast between decision theory and game theory in this regard is less than claimed, a more nuanced methodological perspective would be preferable.

Finally, reading the book left me somewhat confused about how important a role is assigned to social norms and hence to the

choreographer. Social norms are commonly understood to often go against individual benefits. Yet in Gintis' view, social norms function as correlation devices only if they signal strategies that are *best replies* for all players involved (Theorem 7.3). This is certainly not the only way to deal with the interaction of social norms and game theory. Cristina Bicchieri (2006, 3), for example, suggests that social norms transform mixed-motive games into coordination games. But it may be the most conservative one, leaving a large range for decision theory and letting the dancing happen only where its ordnance does not reach.

To conclude, this is an ambitious project, and an exciting one. Gintis draws on many different strands of research, and presents interesting findings that will be new to many social scientists. Yet in order to support his main thesis, he sometimes oversells the confidence we can have in these theories and their results, and he gives short shrift to alternative perspectives that would seem relevant. In a book that essentially argues for the unification of the behavioural sciences, such one-sidedness appears to be a major weakness, as one could suspect that the proposed division of labour is mainly determined by the author's personal predilections. Nevertheless, it is an important and courageous attempt, and a starting call for more research in this direction. May the dance continue!

REFERENCES

Aumann, Robert J. 1976. Agreeing to disagree. *The Annals of Statistics*, 4 (6): 1236-1239.

Bernheim, B. Douglas, and Antonio Rangel. 2009. Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124 (1): 51–104.

Bicchieri, Cristina. 2006. *The grammar of society: the nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Gigerenzer, Gerd, and Henry Brighton. 2009. Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, 1 (1): 107-143.

Hodgson, Geoffrey M. 2008. Prospects for economic sociology. *Philosophy of the Social Sciences*, 38 (1): 133–149.

Mill, John Stuart. 1844 [1836]. On the definition of political economy, and on the method of investigation proper to it. In *Essays on some unsettled questions of political economy*, J. S. Mill. London: Batoche Books, 86-114.

Loewenstein, George. 1999. Experimental economics from the Vantage-Point of behavioural economics. *Economic Journal*, 109 (453): F25-F34.

Vanderschraaf, Peter. 1998. Knowledge, equilibrium and convention. *Erkenntnis*, 49 (3): 337-369.

Till Grüne-Yanoff is a fellow of the Collegium of Advanced Study at the University of Helsinki. Previously, he held appointments at the Royal Institute of Technology, Stockholm, and the London School of Economics. His research focuses on the methodology of economic modelling, on decision and game theory, and on the notion of preference in the social sciences. He has published in *Synthese*, *Erkenntnis*, *Theoria*, *Journal of Economic Methodology*, amongst others, and has edited (together with Sven Ove Hansson) a book on *Modelling preference change* (Springer, 2009).
Contact e-mail: <till.grune@helsinki.fi>
Website: <http://www.mv.helsinki.fi/home/gruneyan/>