

ROUTLEDGE FRONTIERS OF POLITICAL ECONOMY

Economic Indeterminacy

A personal encounter with the
economists' peculiar nemesis

Yanis Varoufakis

2014



Economic Indeterminacy

This volume is a collection of some of the best and most influential work of Yanis Varoufakis. The chapters all address the issue of economic indeterminacy, and the place of a socialized *homo economicus* within the economy. The book addresses Varoufakis' key interpretation regarding the way in which neoclassical economics deals with the twin problems of complexity and indeterminacy. He argues that all neoclassical modelling revolves around three meta-axioms: methodological individualism; methodological instrumentalism; and the methodological imposition of equilibrium.

Each chapter is preceded by an introduction, which explains its place within the overarching theme of the book. The volume also includes a detailed introduction, plus a concluding chapter focusing on the future of economics. It will be a key work for all students and researchers in the fields of political economy and economic methodology.

Yanis Varoufakis is Professor of Economic Theory and Director of the Department of Political Economy within the Faculty of Economic Sciences at the University of Athens, Greece.

Economic Indeterminacy

A personal encounter with the economists' peculiar nemesis

Yanis Varoufakis

 **Routledge**
Taylor & Francis Group

LONDON AND NEW YORK

Contents

List of figures

List of tables

Preface

Acknowledgements

- 1 Introduction: Economic indeterminacy and the dance of the meta-axioms: the dynamic mechanism by which neoclassical economics turns defeat at the hands of indeterminacy into unassailable dominance
- 2 Unity is strength: it is also the cause of indeterminacy regarding the wage and employment preferences of employers and trades unions
- 3 Rational conflict: on the impossibility of a determinate theory of costly disagreement
- 4 No bluffing please, we are economists! Why bluffs and other subversive acts preclude determinate game theoretical analyses
- 5 Bargaining by rules of thumb: when strategic indeterminacy forces the rational negotiator to fall back on myopic rules of thumb
- 6 Marxists and the sirens' song: when disgruntled Marxists reach for neoclassical economics' toolbox they get more than they bargained for
- 7 A theory of solidarity: why indeterminacy is a prerequisite for genuine solidarity
- 8 On the power of what others think: how indeterminacy explodes when our preferences are influenced directly by other people's beliefs
- 9 The social foundations of corruption: on the indeterminate power of what others think
- 10 Evolving morals in the laboratory: the roots of distributional justice principles in indeterminacy
- 11 Evolving domination in the laboratory: the spontaneous creation of hierarchies and the patterned beliefs that support them
- 12 On the distinction between evolution and history: the impossibility of modelling behavioural mutations amongst political animals
- 13 Conclusion: dealing with indeterminacy on the stage of social life

References

Index

Preface

This is a book about failure and power.

Most of us were raised to imagine that power stems from success, not failure. It thus seems odd to be focusing on failure and power, especially when suggesting, as this book strongly does, that massive failure has been the cause of spectacular success. Yet, the world we live in has made possible this sad, odd causality which stands Charles Darwin on his head.

The reader is, at this point, excused to think that the alluded power-through-failure phenomenon refers to the post-2008 spectacle of tremendous taxpayer-funded rewards for deeply insolvent bankers. While this is not my book's theme, the association is not baseless. For, just as the financial sector's implosion yielded its custodians, the bankers, gargantuan rewards (in terms of bailout funding but also of political influence), so too the theoretical failure of mainstream economics has helped solidify and propagate the dominance of these same economists in academia and in the corridors of power.

In this sense, the theme of the present book is very much in tune with our post-2008 age. Yet its origins go back to the 1980s when, as a young, green-behind-the-ears economist, I attempted to build a research programme on several attempts to civilise mainstream economic models that had arrested my attention. It all began at a time when the takeover of economics by a particularly narrow economic method (which I, and many others, refer to as 'neoclassical') had been completed. Those of us who were coming through the academic production line in the UK of the early Thatcher years faced a stark choice: either work within the neoclassical mindset or seek alternative careers. It was that simple.

Determined to master the discipline which in our times represents the highest form of ideology, I was reluctant to abandon economics just because its assumptions and models seemed problematic, if not downright barbaric. In view of the profession's intolerance of any challenge to its neoclassical method, I decided I would attempt two things: to investigate the logical coherence of the received models (i.e. to see if their results were truly consistent with their own assumptions); and to explore ways and means of 'civilising' these models (by relaxing some of their more obnoxious assumptions).

Thus emerged a research project that lasted thirty odd years. Its aim was to add to mainstream models' dimensions (in the form of carefully selected equations) that humanise them, and generally to experiment with their capacity to embrace parts of the social dimension of life that economics had hitherto not even tried to reach. Each of the chapters that follow (after the first, introductory, chapter) revolves around one of these models, telling a story of some attempt to infuse them with a degree of realism, and internal logic, that they lacked.

In retrospect, the research programme which I embarked upon in the early 1980s, and whose models populate the rest of this book, resembled ... invading Russia: a brisk and enthusiastic start, followed by a slowdown as General Winter mounted his hideous counter-attack, ending up with exhaustion, disappointment and metaphorical blood on the snow. Less allegorically, my initial tampering with my new profession's models was met with distinct approval, from professors and editors alike, and job offers that allowed me to claim a place on the academic ladder as a *bona fide* economist. However, from a very early stage, I realised that the profession's welcoming arms would quickly be withdrawn the moment one's model-tampering yielded indeterminacy.

Put simply, while the profession was more than happy to allow newcomers to toy with their assumptions (as the method remained fully neoclassical), it was adamant that models should be 'closed' come-what-may; that our equations should procure a narrow range of 'solutions' even if the only way of achieving such 'closure' was to abandon the

project of civilising the theory. As far as the economics profession was concerned, logical incoherence and a deep chasm between the models and really-existing capitalism were infinitely preferable to an admission that the models were indeterminate.

At first, my peers' profoundly anti-scientific attitude disturbed me no end. Until, that is, it all started making sense in a broader political economics context. To begin with, I noticed an interesting paradox develop from the time I was an undergraduate in the late 1970s: the more dominant economics was becoming within academic social science the more students were being turned off economics. Instead of magnetising the young, courtesy of its indubitable discursive success, economics was putting them off.

And it was not just students. Economists of renown were lambasting their discipline's irrelevance and theoretical feebleness. Nevertheless, and there is the rub, *the greater the mainstream economists' theoretical failure the stronger their dominance everywhere*. How come? A major clue to this puzzle came in the form of the observation that these same models, precisely because they turned a blind eye to the indeterminacy that oozed out of them, were also the models underpinning the financial derivatives that the financial sector was beginning to invent at that time (which it soon flooded the world of finance with), as well as the neoliberal doctrines which were used as a pretext for engineering the most regressive income redistribution in the history of capitalism.

Faced with this disturbing, but also deliciously ironic, reality I chose to tread a thorny path: I would continue to tamper with the mainstream models' assumptions in a bid to explore their explanatory potential to the full. At the same time, I

would expose the logical contradictions of the models that my profession deemed beyond analytical reproach. And, lastly, I would attempt to provide an explanation of the manner in which neoclassical economics was building impressive discursive power on a foundation of large-scale theoretical failures.

Naturally, my project's failure was predetermined, at least in the sense that it was never going to cause a shift in the attitudes and demeanour of a profession which operates like a priesthood, dedicated solely to the preservation of its dogmas (which I call meta-axioms in

Chapter 1

) as well as to the recapitulation of its authority within the universities, the financial sector and government. Indeed, at no point did I harbour any significant hope that this priesthood would take kindly to the demons of doubt and indeterminacy which my work was bound to give rise to. But it did not matter, at least not at a personal level. My intimate familiarity with the neoclassical models was sufficient to keep me on the roster of neoclassical economics departments, where a capacity to teach these models, and produce academic papers based on them, is all that matters.

Looking back at these long years of tampering with, and delving into, the complex models of the neoclassical tradition, I cannot but question my decision to keep pushing, Sisyphus-like, the theoretical rock up the neoclassical hill. Why did I stick to this task, when I knew it would end up in failure? In retrospect, there were two reasons, neither of which was predicated upon any hope of influencing a profession utterly uninterested in the truth-status of its models. First, I deeply enjoyed toying with these models as an end-in-itself, just as a clockmaker enjoys taking apart and then re-assembling some old clock for the hell of it. Secondly, and more importantly, I felt it necessary and important to make it hard for my colleagues to pretend to themselves that the models they were being forced to work with, by a particularly authoritarian profession, were logically coherent. Bringing them, even fleetingly, to the point when they *had* to confess to their models' internal contradictions was, I felt, a victory of sorts; the equivalent of a lone sniper behind enemy lines making life difficult for an army of occupation.

At the end of the day, I now realise that failure is indeed packed with power, not just

for bankers and the economics profession but for us mere mortals too.

Chapters 2

to

11

, in effect, explore theoretical failures. Indeed, while working on these models I often caught myself at the intersection of many failures: mathematical, philosophical, conceptual. However, coming to terms with these failures was essential in understanding the irrationality of the world we live in. For these failures are not the result of substandard skills or erroneous manipulations but, rather, a mere reflection of the dead-end forced upon us by an ideologically driven pseudo-science whose power comes from successfully hiding, as opposed to revealing, the true nature of our social, political and economic relations.

Acknowledgements

The story of how neoclassical economics profits from its failure to deal with complexity, with radical indeterminacy beckoning at each and every turn, was told fully in a recent book co-authored by myself, Joseph Halevi and Nicholas Theocarakis (*Modern Political Economics: Making sense of the post-2008 world*, also published by Routledge in 2011). It is, therefore, incumbent upon me to begin by thanking Joseph and Nicholas for having helped in creating the context which the present book builds upon. Indeed, the book you are now holding can be seen as a natural addendum to *Modern Political Economics*, as the following chapters demonstrate, one by one, the manner in which neoclassical economics turns failure into power once all attempts to civilise its models yield radical, irrepressible indeterminacy (an adage that our joint book foreshadowed; see its

[Chapters 9](#)

and

[10](#)

in particular). Christian Arnsperger, another friend and colleague, has been responsible for the impetus behind

[Chapter 1](#)

, as well as being the co-author of a paper on which

[Chapter 7](#)

is based. Shaun Hargreaves-Heap played a major part in joint work on which I have based

[Chapters 8](#)

and

[11](#)

. Lastly, I must thank the good people at Routledge, Simon Holt and Robert Langham in particular, for bearing with me during long delays in the production of this manuscript, which I blame on the European Crisis that has kept me 'otherwise engaged' for too long.

1 Introduction: Economic indeterminacy and the dance of the meta-axioms

The dynamic mechanism by which neoclassical economics turns defeat at the hands of indeterminacy into unassailable dominance

1.1 Prologue

Since the 1970s, give and take a few years, a decision to immerse oneself in economics has been translating into an exclusive training in what can be termed *neoclassical economics*. Neoclassical economics is a particularly narrow method of conceptualising market economies which, astonishingly, has managed to monopolise academic and professional economics since the mid-1970s. In philosophy, this would be the equivalent to, say, a total supremacy of Existentialism over all other philosophical traditions – to the extent that Plato, Aristotle, Hegel and Russell do not even get a mention in any of the offered courses. Such a development would be scandalous and impossible to imagine. And yet, in economics this is our reality.

How did the bewildering dominance of neoclassical economics come to pass? By what mechanism is it reproduced? In this chapter I venture an answer to these two questions.

Section 1.2

defines what I mean by neoclassical economics; the sort of economic analysis that succeeded in expelling all other analytical methods from professional economics. My definition is that all neoclassical modelling revolves around three meta-axioms: methodological individualism, methodological instrumentalism and methodological equilibration. Then,

Section 1.3

presents the dynamic mechanism, which I call the *dance of the meta-axioms*, by which neoclassical economics reinforces its power. Central to this process is, I argue, the way in which neoclassical economics deals with the twin problems of *complexity* and *indeterminacy*.

To deliver models of reasonable complexity, neoclassicism relaxes the first two meta-axioms. However, the price of that relaxation is massive, radical indeterminacy. To arrest it, neoclassicists resort to an austere tightening of the third meta-axiom. The *dance of the meta-axioms* is, therefore, a series of unending moves by the economics profession: When interested in demonstrating the sophisticated complexity of their models, they take forward steps through the relaxation of meta-axioms 1 and 2. But when indeterminacy threatens to dissolve the offered analysis, they move sideways and then backwards through the tightening of meta-axiom 3. Soon after, however, in a bid to return to a sophisticated narrative, they relax, once more, meta-axioms 1 and 2. And so on.

Section 1.3

concludes with a number of examples of such moves in the literature; moves that have shaped the profession's views on all the significant theoretical issues that economists are naturally interested in – from the theory of value and growth to game theory and the theory of risky choices.

Section 1.4

discusses how the *dance of the meta-axioms* helps reproduce neoclassicism's dominance, by extracting copious discursive power from its spectacular theoretical failures. It tells a story of how young, gifted economists are lured into the worst type of theoretical cynicism, lashed with generous doses of solipsism. Almost in a bid to

emulate the bankers' capacity to extract huge rents from society after the financial crash of 2008, i.e. to benefit in proportion to their failures, so too neoclassical economics succeeds in reinforcing its dominion in proportion to the magnitude of its theoretical failure. A most peculiar 'failure,' indeed...

So, the *dance of the meta-axioms* is central to this book. Every chapter that follows represents a personal encounter with this dance. It constitutes a case study of how a reasonable attempt to relax the more unrealistic assumptions of certain models that I once studied ended up in retreat and ignominy due to the profession's determination to 'close' the models down; to dance the *dance of the meta-axioms*. Taken together, these chapters comprise a warning for graduate economics students and young academic economists that alerts them to the unintended ruthlessness with which they will be met, by the profession, if they put realism and intellectual honesty above the urge to 'close' their models, even if this is what common sense demands of them.

In summary, the remainder of this chapter argues that:

- (a) neoclassical economics is well defined in terms of three meta-axioms (*methodological individualism*, *methodological instrumentalism*, and *methodological equilibration*);
- (b) their adoption is the common practice which delineates mainstream economics;
- (c) while the first two meta-axioms allow for rich depictions of socioeconomic phenomena they lead to an unquenchable indeterminacy, and
- (d) the spectre of this indeterminacy generates evolutionary and social forces within the economics profession which cause practitioners to introduce stringent variants of the third meta-axiom.

Thus the neoclassical models' sophisticated complexity is sacrificed in favour of a determinate framework within which not even a glimpse of contemporary capitalism is possible. Neoclassicism, I contend, owes its hegemonic position in the social sciences to this most peculiar, axiomatically inbuilt, theoretical failure that is masked and turned into stunning success by a series of moves I term *the dance of the meta-axioms*.

1.2 The three meta-axioms underpinning neoclassical economics

Few, if any, economists would describe their work as neoclassical. As the term was coined much later, the nineteenth century pioneers of marginalism would not have even recognised it. As for contemporary economists, they seem ill disposed to the neoclassical label even when their work is demonstrably neoclassical.

¹ But this disinclination, in itself, is immaterial: for if a particular body of economics can be *profitably* distinguished by means of some single epithet (e.g. 'neoclassical'), the deployment of such an epithet may be in order. After all, the inhabitants of the Eastern Roman Empire would not have appreciated the label 'Byzantine'; nor would late nineteenth century Britons have conceived of their society as 'Victorian.' Such epithets have analytical value analogous to their capacity to illuminate certain eras and mind frames.

In my quest for a useful definition, I take a second leaf out of the historians' book: Their terms 'Byzantine' or 'Victorian' may well be over-arching but, at the same time, are deployed carefully so that their use does not invalidate their subject-matter's *dynamic complexity*.

² In the same vein, we too ought to be keen to define neoclassical economics in a manner that respects the undisputed fact that its axioms and theoretical practices have been evolving, changing, and adapting from the very beginning. For that reason, I shall eschew any definition based on a fixed set of neoclassical axioms.

Let's begin by asking: Granted that neoclassicists' axioms and methods are in constant flux (inter-temporally but also across different models and fields), is there some analytical foundation which: (a) remains time and model invariant, and (b) typifies a distinct approach to economics? This is equivalent to searching for invariant *meta*-axioms: higher-order axioms about axioms which underpin *all* of neoclassical economics, irrespective of the actual axiom's fluidity or the malleability of its focus. I propose three such meta-axioms as the foundation of all neoclassicism.

1.2.1 Meta-axiom 1: methodological individualism

Consider the analytic-synthetic method of a watchmaker faced with a strange mechanical watch. First, she takes it carefully apart with a view to examining the properties and function of each of its tiny cogs and wheels. Then, she screws it back together. If a reassuring ticking sound ensues, this must surely mean that the fragments of knowledge imparted by the *separate* study of *each* of its parts were successfully synthesised into a macro-theory of the watch.

This parable of an ideal reductionist, analytic-synthetic economic approach has been implicit to neoclassical theorising since the first stirrings of marginalism. While the term *methodological individualism* came later with Schumpeter (1908), it featured well before its christening as the bedrock on which economics (in juxtaposition to classical political economy) was to be re-founded. To the economists who sought a break from the political economy of Smith, Ricardo and Marx, a new focus on the individual agent became the litmus test of 'scientific' economics (see Mirowski, 1989).

In this new, or neoclassical, mindframe, individuals are the equivalent of the watchmaker's cogs and wheels: parts of a whole to be understood *fully* (complete with determinate behavioural models) and *independently* of the whole their actions

help bring about. Thus, any socio-economic phenomenon under scrutiny is to be explained *via* a synthesis of partial knowledge derived at that individual level.

But there is a snag: Unlike the world of mechanical watches, society consists of 'parts' which are not readily separable. A pulley or a cog can be fully described in isolation to the other mechanical parts with which it was designed to work harmoniously. Indeed, the 'relations' between the watch's parts are straightforwardly revealed, to the trained eye, through close inspection of the parts' shape, size and other physical properties. In the social world, however, not only are the relations between its 'parts' not deducible from primitive data concerning these parts alone (e.g. from data on persons' means and ends) but also it is simply impossible to understand the parts' properties in isolation to one another. When Aristotle spoke of humans as political animals, or when Hegel narrated his master-slave paradox, they were dwelling on this radical difference between the constituents of society as opposed to the parts of mechanical systems (regardless of their complexity).

Hodgson (2007), drawing on Udéhn (2001, 2002), relates the ambiguities in the methodological individualism espoused by leading neoclassicists and suggests that neoclassicism seems to oscillate between strong methodological individualism, which insists that all explanation must be reducible to knowledge derived from isolated selves (an archipelago of Robinson Crusoes), and a weaker version which acknowledges that the individual is indefinable outside its social and relational context. My explanation of this oscillation will be that, while thoughtful neoclassicists are mindful of the logical conundrum awaiting them if the analysis of persons excludes their relations to other persons (and, thus, to the surrounding institutions), they are *forced* inevitably to fall back on a strong version of methodological individualism.

Forced by what? By the ambition to 'close' their models, I suggest (see Lawson, 2003, for the predilection of mainstream economics for closed explanatory systems). Human relations are notorious for their resistance to determinate modelling. Put simply, the mathematics of defining a person in terms of her relations to others, in addition to

her means and ends, is of an order higher than most economists would want to engage with and, worse, offer no determinate solution (i.e. behavioural prediction).

4

Importantly, this is no mere technical difficulty awaiting a technical fix. Rather, it reflects the *impossibility* of a deductive methodological individualism which treats human relations as primitive data (see also Fine, 2008). It is for this reason that neoclassicism gravitates toward strong methodological individualism, while alluding to its weaker version when in a more philosophical mood.

To sum up, neoclassicism's first meta-axiom encompasses two main variants of methodological individualism, one of which typifies neoclassical economics of *all* types: *Strong methodological individualism* – **D**: All explanations are to be synthesised from separate, autonomous, and prior explanations at the level of the individual. A *strict explanatory separation* of *structure* from *agency* is imposed, with an analytical trajectory that moves unidirectionally from full explanations of agency to derivative theories of structure. In this variant, agency feeds into structure (which is merely the crystallisation of agents' past acts) with no feedback effects from structure back into agency.

Weak methodological individualism – **d**: As above, with the difference that feedback between structure and agency is permitted, even though the explanatory force remains in the realm of agency.

All textbook economics is founded on **D**, as are the foundational texts on the mainstream's main theorems: general equilibrium, game theory, new classical economics etc. However, in the last two decades or so, a new crop of highly interesting models has appeared which turn on **d**.

5

In the following sections I shall be arguing that the interplay between **D** and **d**, rather than signifying a retreat from neoclassicism, is part of a complicated dynamic which reinforces its dominance and can be grasped only when all three meta-axioms are considered at once. Therefore, I now turn to the other two meta-axioms.

1.2.2 Meta-axiom 2: methodological instrumentalism

Methodological individualism is vacuous without a theory of what motivates individuals. Contrary to the impression given by microeconomics textbooks, greed was never a foundational assumption of neoclassicism. While it is true that its models may have been traditionally populated by hyper-rational bargain-hunters, never able to resist an act which brings them the tiniest increase in expected net utility, the latter can just as readily result from bars of gold as from reductions in third world poverty.

Closer to the truth, regarding neoclassicism's foundations, is the claim that it relies on the axiom of *instrumental* (or means-end) *rationality*: Agents are rational to the extent that they deploy their means efficiently in the service of *current, prespecified* and *sovereign ends*. However, I have already explained why I shun any definition of neoclassical economics which turns on some specific axiom. By the term *methodological instrumentalism* I signify a meta-axiom which encompasses all strands of motivation within neoclassical economics (from Jevons and Marshall to evolutionary game theory

6

).

Strict methodological instrumentalism – **S**: Behaviour is driven by some well defined function mapping the combination of all feasible agents' behaviours to some homogeneous index of individuated 'success'. The latter reflects agents' preferences which are given, current, fully determining, and strictly separable both from: (a) belief

7

(which helps the agent evaluate the alternative future outcomes), and (b) the means employed.

Weak methodological instrumentalism – **s**: Behaviour is, again, explained in terms of a homogeneous index of ‘success’, onto which behaviours are mapped. However, the focus of study is no longer the decision maker but rather each element of her complete set of feasible actions (*aka* strategies). The models are, in this sense, populated by competing alternative strategies or behaviours (rather than decision makers) whose fortunes are determined not by instrumental rationality but by some ‘replicator dynamic’; that is, by a difference or differential equation which ‘selects’ the strategy or behaviour that ‘does better’ than its ‘competitors’ in terms of some exogenously given set of individual ‘welfare’ criteria.

8

Under both **S** and **s**, rationality loses its *substantive* meaning. **S** turns rationality into a capacity to achieve the highest possible level of preference-satisfaction, so much so that there is no longer any philosophical room for questioning whether the agent will/should act on her preferences.

9

Bounded ‘rationality’ is also permitted, under both **S** and **s**, when the computation of optimal decisions is costly and/or time consuming. Lastly, under **s**, substantive rationality is wholly absent (since humans are not even the object of study in these models) and yet the analysis is fully instrumental as behaviour is selected (or abandoned) on the basis of fully specified exogenous goals.

10

Before proceeding to neoclassicism’s final meta-axiom, it may be of interest to note that both strands above, **S** and **s**, can be traced to David Hume (1739/40, 1888). The origins of **S** lay in his famous division of the human decision-making process into three *distinct* modules: *Passions*, *Belief* and (instrumental) *Reason*. *Passions* provide the destination while *Reason* slavishly steers a course that attempts to get us there, drawing upon a given set of *Beliefs* regarding the external constraints and the likely consequences of alternative actions.

11

As for **s**, and neoclassicism’s ‘evolutionary turn’, it too draws its energy from the *Treatise* and in particular from the argument that, when instrumental reason is given insufficient ‘data’ on which to base a firm decision (a case of ‘multiple equilibria’, in today’s parlance), conventions or customs emerge that fill in the vacuum. Their evolution proceeds along the lines of an adaptation mechanism which selects practices according to their efficacy, viz. the agents’ pre-determined passions.

12

Where **s** diverges sharply from Hume is in its incompatibility with the one thing he cared greatly about: the (unmodellable) feedback effect between, on the one hand, forecast, action and, outcome and, on the other, the normative beliefs that are born endogenously

13

and which fashion our view of that which we call our ‘self-interest’.

1.2.3 Meta-axiom 3: methodological equilibration

All economics revolves around the search for equilibrium states or paths, ranging from the theories of Ricardo, Marx and Sraffa to the neoclassicists.

14

What distinguishes neoclassicism, in this regard, is that equilibration is usually imposed axiomatically even in the absence of any plausible explanation of how the system under study is supposed to edge closer to equilibrium. This practice is best described as a meta-axiom since it takes many different axiomatic forms which, nonetheless, are consistent with the definition of strong methodological equilibration below:

Strong methodological equilibration – **E**: Once the set of equilibria is deduced from the available primitive data (e.g. motivation, constraints, production possibilities, adaptation

mechanisms, etc.), the focus of study is restricted (usually by some hidden axiom) to that set and only behaviour consistent with it is admitted. Sensitivity analysis is then introduced to discern the equilibria at which small, random perturbations are incapable of creating centrifugal forces able to dislodge behaviour from that state or path.

15

Weak methodological equilibration – **e**: The set of equilibria is arrived at through a process that unfolds either in logical or historical time by means of a pre-specified selection mechanism which forms part of the analysis' primitive data.

The classical economists, also beholden to equilibration, traditionally espoused **e**.

16

Pre-1950 neoclassical models also refrained from **E**, investing their skills in devising logical explanations of the path to equilibrium.

17

However, the slide from **e** to **E** began in earnest first with John Nash's approach (1950, 1951), to the bargaining problem in particular and to strategic action in general, and then with Debreu and Arrow (see Debreu, 1959, and Arrow and Debreu, 1954) who, following a presentation by Nash at the Cowles Commission in October 1950, abandoned **e** in favour of **E**.

18

The outcome of this radical shift was the celebrated proof of the existence of general equilibrium prices; a proof purchased at the cost of historical time (and, thus, of any logical argument regarding how that general equilibrium might emerge in time).

19

1.3 The *dance of the meta-axioms*

Models are an open invitation to meddle with assumptions, and neoclassical models have been no exception. After several decades of such meddling, and with new empirical and computational techniques increasingly being pressed into service, many economists, including some who have been critical of the mainstream,

20

began to discern a fundamental shift from neoclassical formalism toward a new methodological pluralism. In evidence, they cite the noteworthy makeover that *homo economicus* seems to have undergone

21

and, more generally, the observation that the traditional neoclassical core (e.g. general equilibrium, the neoclassical macroeconomics synthesis) seems eclipsed – immersed – in the shadows of game theory, nonlinear models, experimental economics, simulations, neuroeconomics, evolutionary models etc.

This section cautions against such a conclusion. It suggests that, on close inspection, the centrifugal forces occasioned by dissatisfaction with the original formalist neoclassical position, after initially pushing the mainstream away from the neoclassical nucleus, eventually subside, turning centripetal. Thus, they return the offered analysis either to the original neoclassical position or, even worse, to a position at a higher plane of neoclassical abstraction on which the original 'problem' not only remains unsolved but is, indeed, *amplified*.

The dynamic mechanism at work is outlined in

Figure 1.1

in diagrammatic form. I refer to it as the *dance of the meta-axioms* featuring the following simple

'steps': Starting from **1**, the *original formalist neoclassical position*, some theoretical *challenge* is issued (either from within neoclassicism or from without). In some cases, the challenge is *ignored* outright (arrow **i**) while in others it is *addressed* (arrow **a**) *via* a relaxation that occurs within one or both of the first two meta-axioms. At that stage, we argue, radical indeterminacy sets in and the profession recoils: Either it *retreats* to the original position (**1**) or it *backslides* (arrow **b**), *via* a severe tightening

of the third meta-axiom, to some new position **4**; a position where the original problem (that **c** sought to address) seems assuaged when, in truth, its intractability is greatly intensified.

The remainder of this section illustrates this hypothesised dynamic by evoking a number of challenges (**c**) to core neoclassical models and groups them under our three main trajectories. I begin with important challenges which were ignored outright (**i**). Next, I turn to challenges of note which were addressed (**a**). From some, the profession retreated (**r**) while others occasioned a backslide (**b**) to a new, more complex neoclassical position even more theoretically problematic (but also discursively more powerful) than the original.

Essential to my hypothesis is the argument that: (*i*) none of these challenges could penetrate the resulting wall of indeterminacy while retaining their allegiance to the neoclassical meta-axioms, and (*ii*) the profession, after dallying with complications of its foundational neoclassical models, returns to a position (**1** or **4**) which, at the expense of explanatory power, remains as contained within the meta-axioms as ever.

1.3.1 Ignored challenges: the 1→2→1 quickstep

In this subsection I look at challenges to the neoclassical method which, while poignant and valid, were unceremoniously ignored by the mainstream. I begin with the 1950s explosion of neoclassical decision and game theory that was founded on expected utility theory (as outlined by von Neumann and Morgenstern, 1944; and Savage, 1954). From a very early stage, its foundational assumptions were challenged both experimentally and logically. In particular, two separate but equally devastating critiques, by Allais (1953) and Ellsberg (1956, 1961), disproved the empirical validity of expected utility theory and challenged the logic of its foundational axioms. Since then a cottage industry of laboratory experiments has confirmed the former while a series of fascinating alternatives to expected utility theory have been published in the mainstream's top journals (for surveys see Sugden, 1991; and Starmer, 2000). And yet, to this day, expected utility theory reigns supreme both in the lecture theatres and in every form of neoclassical theorising, from rational expectations models to each and every application of game theory.

In game theory itself, questions were raised about the plausibility of presuming that rational agents must always select behaviour consistent with Nash's (1951) equilibrium. In the context of static games it became apparent that *disequilibrium* behaviour could be fully rationalised and rendered consistent with infinite order common knowledge rationality.

²²

Similarly, it transpired that out-of-equilibrium

behaviour could be just as rational in finite dynamic games as the equilibrium path proposed by Nash and his disciples.

²³

As for indefinite horizon games, the devastating force of indeterminacy was felt in the form of the so-called *Folk Theorem* which shows that, in interactions that last for an unspecified period, *anything goes*.

²⁴

And yet, *all* applications of game theory, from theories of Central Bank behaviour to industrial organisation, labour economics and voting models, ignore these challenges, assuming that behaviour will remain on *the* equilibrium path.

²⁵

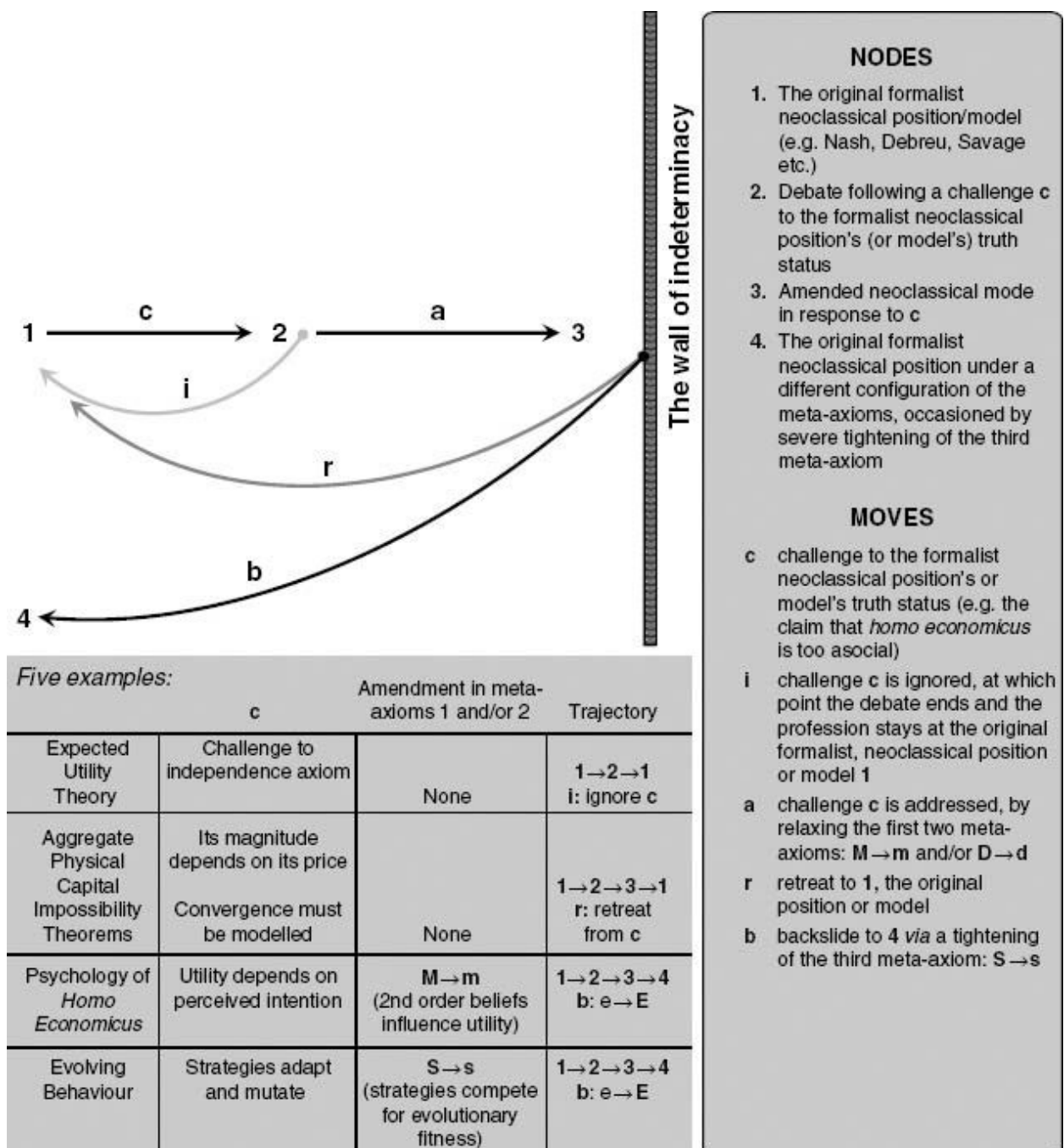


Figure 1.1

The dance of the meta-axioms.

Perhaps the best known case of a challenge ignored is the debate known as the *capital controversies*. Neoclassicism traditionally insisted that, with price taking agents, returns to capital reflect capital's marginal productivity. The challenge to

this notion came from Cambridge economists Piero Sraffa, Joan Robinson and Luigi Pasinetti who pointed out a highly damaging reflexivity: While it is possible to speak meaningfully of homogeneous apple juice, even of homogenous 'abstract' labour, it is impossible to treat capital goods as homogenous (in view of their different types and vintages) and, consequently, to measure an economy's capital stock independently of its price. But then, if physical capital's magnitude depends on its price, how can its price be explained by its magnitude? This challenge prompted a series of exchanges (see Harcourt, 1972) which petered out once the neoclassical corner effectively threw in the towel.

And yet, today, no trace of this debate is to be found in any mainstream economics

curriculum. The challenge has been ignored and the mainstream has continued to assume that the profit rate (i.e. capital's price) is explained, uni-directionally, by the revenues due to the last morsel of an aggregate physical capital whose magnitude is independent of that return. All the developments of the 1970s and beyond (rational expectations, new classical and recursive macroeconomics, etc.) proceeded as if this debate had never taken place.

27

1.3.2 Retreat: the 1→2→3→1 move

Not all valid and poignant challenges came from critics of neoclassicism. Some of the strongest ones emanated endogenously and, perhaps for this reason, were taken seriously by the profession. The best example relates to the theorem by which a general equilibrium was proven to exist: Arrow and Debreu (1954) arrived at their celebrated proof by first taking a leaf out of Nash's proof of the existence of a unique solution to the bargaining problem (see Nash 1950). The key idea that they borrowed from Nash was *to abstract fully from the equilibration process*.

28

Adopting the strong version (E) of the third meta-axiom, Nash and Arrow and Debreu established their unique equilibria only *by purposefully ignoring the movements leading to it*. The profession was, understandably, dazzled by these remarkable existence proofs. Nevertheless, it was not too long before questions were being asked about how the equilibrium obtains in real time (either in bargaining or in some multi-sector neoclassical economy).

While Nash and Debreu had no qualms in admitting that it was part of their proof *not* to have an answer to this,

29

neoclassicism could not avoid such questions, especially in the lecture theatres. Teachers found themselves almost compelled to rely on deeply unsatisfactory heuristics. In the case of bargaining, stories were told that involved positing a bargaining process with stages in which concessions were motivated by different amounts of fear of disagreement.

30

Similarly, in the case of the competitive price mechanism, tales of equilibration were allowed to linger on the basis of an analytically untested belief that prices must adjust until excess demand vanishes.

While these equilibration narratives had (and could have had) no basis in the axiomatics of Nash or Arrow–Debreu, they *seemed* ever so obviously correct to students as to silence all dissenting voices. Except, of course, those of the *leading* neo-classicists, who understood only too well the analytical folly intrinsic

to these. Nevertheless, with one exception (namely, Debreu

31

) they craved some demonstration of convergence to their axiomatically derived, and thus inherently static, equilibria; a demonstration with which to replace the incongruous lecture theatre tales. Thus, a challenge was issued, from *within* neoclassicism, to model convergence explicitly, both in the context of general equilibrium and in bargaining. Indeed, in a world of disequilibrium, flux, persistent unemployment, periodic price wars, painful industrial disputes etc. – an inability to say something meaningful on out-of-equilibrium prices, or on costly delay before reaching agreements – would have been tantamount to a declaration of theoretical failure.

In the case of Nash's bargaining theorem, Rubinstein (1982) rose to the challenge: Nash's solution, he argued, could be shown to be the limiting case of a bargaining process in which rational bargainers issued alternating demands.

32

As for costly delays in reaching agreement, they could be explained by asymmetrical information on each other's eagerness to settle (see Rubinstein, 1985). In

general equilibrium theory, some promising preliminary work hinted at ways in which the groping process toward an equilibrium price vector could be modelled (for an early attempt see Arrow, 1959). However, it was not long before it transpired that both projects were doomed. The bad news for the neoclassical project, in both cases, came from Hugo Sonnenschein and his collaborators.

Starting with general equilibrium, Sonnenschein (1972, 1973) demonstrated (confirming Debreu's stance – see note

³¹

) that excess demand for some commodity could never guarantee that its price would rise; that, even if individual excess demand functions were well defined, aggregate demand was not. The implication was startling and its poignancy confirmed by Mantel (1974) and Debreu (1974). The combined meaning of what has become known as the *Sonnenschein–Mantel–Debreu theorem* (SMD hereafter) was: (a) that convergence to general equilibrium is *impossible* to model, and (b) that it is no longer possible to guarantee the general equilibrium's uniqueness. Moving on to bargaining theory, the idea that delay in reaching agreement could be explained by asymmetrical information, within the context of the Nash–Rubinstein approach to bargaining, was dispelled by Gul and Sonnenschein (1988). In conjunction with the aforementioned devastating critique of the logical coherence of subgame perfection in dynamic contexts,

³³

the literature reached a simple conclusion: Rationality (of whatever order, breadth, extent or commonality) could *never* ensure that a bargaining process between rational agents is amenable to the mathematical modelling of some stochastic equilibrium path.

³⁴

Taken together, these two contributions had a single, inescapable, implication for the grand neoclassical project of the 1950s: *The highest form of neoclassicism had nothing meaningful to say about price and contract formation*. Intriguingly, it was neoclassicism which challenged itself to come up with a response to the convergence issue and it was neoclassicism which procured these two impossibility theorems which *prove* that it could not meet its own challenges. In terms of the previous section's meta-axioms, the point here is that the best and brightest challenged themselves to shift the highest form of neoclassical theory away from a

reliance on version **E** of the third meta-axiom and toward its weaker version **e**. Alas, all such efforts crashed against a wall of indeterminacy.

The crucial question is: What happened next? The answer is: A multifarious retreat (arrow **r**) back to position **1** in our diagram! Just as in the case of the Cambridge controversies, the challenge came to naught, even if it was an endogenously generated one which neoclassicists valiantly tried to rise to. The actual retreat (arrow **r**) took various forms. Most common is the retreat behind single sector or representative agent models in which the weak third meta-axiom (**e**) suffices. What is, however, of great interest is the *repeated* deployment of the **1→2→3→1** move: When facing questions about the determination of value in a world of many agents and sectors, the profession responds by showcasing the original Nash–Debreu–Arrow analysis, complete with the strong version of the third meta-axiom (**E**). If fresh questions follow regarding convergence, dynamics, growth etc., the weaker version (**e**) comes into play and the emphasis shifts silently from Nash–Debreu–Arrow to representative agent and/or single sector models. And if anyone, at this point, impertinently protests that the world comprises multiple agents and sectors, her neoclassical interlocutor dusts off Nash–Debreu–Arrow once more and brings on **E**. And so on.

This continual *move* back and forth between **e** and **E** keeps out of sight the theoretical failure to rise to the original challenge **c**. In fact, which of the two versions of the third meta-axiom is deployed depends on the question the neo-classicist feels compelled to answer: If she is put on the spot to explain action (e.g. moves, offers) in

real time, she will deploy **e**. But if she needs to articulate a theory of prices (competitive or bargained, e.g. in neoclassical macroeconomics, labour economics, industrial organisation), she returns to **E** and the glittery existence proofs founded upon it. Above all, the surreptitious, never-ending move from **e** to **E** to **e** to ... *ad infinitum* keeps out of sight the neoclassical failure to rise to its own challenges, and thus out of the mainstream economists' agenda.

Dow (1995) correctly writes that, in juxtaposition to the Keynesian method, which she favours, 'mainstream methodology limits economic theory to those elements of the economic process which can, in practice, be represented by a closed, formal system.' However, when adding that 'a high degree of certainty can be achieved within those limits', she is conceding too much. As we have shown above (and in the next subsection), the said 'certainty' is attained only by logically illegitimate moves involving the covert re-switching between the strong and the weak versions of the third meta-axiom.

1.3.3 Backslide: the 1→2→3→4 shuffle

This subsection discusses two examples of what I call the *backslide* (arrow **b** in the diagram) which, following a failed foray into greater plausibility and sophistication, returns the theory not to its original position (node **1**) but to a state once removed from it (node **4**) where the original position's weaknesses are both *better hidden* and *much amplified*. Our two examples concern, first, the attempts to give *homo economicus* a (much needed) richer psychology and, secondly, neoclassicism's so-called *evolutionary turn*.

Let us begin with the major breakthrough in economic psychology marked by two classic papers: Geanakoplos *et al.* (1989) and Rabin (1993). Jill is now psychologically sophisticated in her interactions with Jack and cares not only about what he will do but also about his motives. To illustrate, suppose that, in a static prisoner's dilemma, and under commonly known rationality (CKR), Jill predicts that Jack will defect. In standard neoclassical analysis, there is nothing more to say: They will both defect and their payoffs will be those that correspond to mutual defection. However, in the psychologically enhanced version, intentions matter. Consider two different thoughts which might be underlying Jill's prediction that Jack will defect:

(A) 'Jack is defecting *because* he is expecting me to defect too'

(B) 'Jack is defecting *even though* he is expecting me to cooperate.'

The point here is that Jill may have a legitimate reason to feel worse under (B).

35

For under (B) she thinks that, by defecting himself Jack is shunning a 'kind gesture' of hers.

36

By contrast, under (A) his defection is deemed psychologically neutral.

37

The analytical significance of the above is (a) that it enhances the analysis' realism (by restoring the motivational role of perceived intentions) and (b) that it allows us to rationalise cooperative outcomes unceremoniously dismissed by standard neoclassical theory.

38

There are two morals to this story (for more see

[Chapter 8](#)

). First, neoclassicists are right when arguing that *homo economicus* can be 'trained' better to resemble a real person through a relaxation of their first two meta-axioms.

39

The second moral, however, is more sobering: Indeterminacy kicks in with a vengeance, causing a backslide to an even less defensible position than the original. The reason is that the attempt to civilise the neoclassical agent threatened to wreck the

very fabric of the analytical framework. What I term 'the backslide' is merely a reaction to this threat.

To see this, note that the standard analysis (featuring psychologically unsophisticated players) requires no more than the weak version (**e**) of the third meta-axiom to yield a unique equilibrium.

In contrast, Jill's and Jack's newfound complicated psychology gave rise to a novel, and particularly, sinister type of indeterminacy: the prisoner's dilemma ceases to be a well defined game!

Indeed, when motives 'infect' utilities directly, the only way of writing down the game's payoffs is if we know the players' beliefs a priori. But we can only know them a priori if we make the a priori assumption that their (first- and second-order) beliefs are aligned! Therefore, to help retrieve the prisoner's dilemma as a well-defined game (i.e. to be able to specify the utilities from each of the interaction's four potential outcomes

), the hapless theorist is forced to backslide to the strongest imaginable version of the third meta-axiom. To an **E** on ... steroids.

The above illustrates nicely the backslide (**b**) in the preceding diagram: A fascinating challenge (**c**), emanating from another field (psychology, in this example),

was taken on gallantly by the profession (arrow **a**) but the ensuing indeterminacy defeated its best intentions and forced it onto the back foot. The indeterminacy proved so radical that it jeopardised not merely the model's 'closure' (i.e. whether a unique solution can be found) but, indeed, the model's very structural coherence.

A major tightening of the third meta-axiom saved the day, *via* a logically indefensible leap of faith,

returning the analysis not to its original position (**1**) but to another position (**4**) once removed from it. Interestingly, at that new position (**4**), the theory is rationally less defensible than before, but simultaneously possesses more discursive power!

The evolutionary turn of neoclassical economics is my second example of a major *backslide*. Evolutionary biologists

demonstrated that, in a hypothesised world of insects and birds, behaviour converges automatically onto neoclassical equilibria; seemingly with no need for the third meta-axiom. Understandably, the mainstream was thrilled by this discovery which vindicated neoclassicism,

sharpened its predictions,

and allowed the deployment of the weak version of its third meta-axiom

on the basis of an intuitively appealing Darwinian rationale. For a moment, neoclassicism's triumph seemed complete; even critics of the mainstream came to see the evolutionary turn as a sign that the mainstream was no longer neoclassical.

Were matters allowed to rest there, the inevitable conclusion would have been that the neoclassical mainstream had been on the right track all along (regarding the substance of its hunches) and that, following its evolutionary turn, it reached a stage of development at which it could afford to stop being neoclassical (that is, to drop the third meta-axiom's strong version) and evolve itself into a quasi-Darwinian, technical albeit pluralist, complexity-friendly and, ultimately, more scientific socio-economic discipline.

Alas, matters could not rest there. For, on closer inspection, it soon becomes clear that the Darwinian mechanism at the heart of neoclassicism's evolutionary turn is methodologically equivalent to the third meta-axiom and a brake on any substantive venture beyond the neoclassical meta-axiomatic straitjacket.

Recall that all evolutionary models turn on two mechanisms: an *adaptation mechanism*, which is responsible for convergence *via* some type of natural selection (or *replicator dynamic*), and a *mutation-generating mechanism* which produces a constant inflow of variety. The aforementioned evolutionary dynamic is based on a joint assumption: (A) that the two mechanisms are independent of each other, and (B) that mutations are identically and independently distributed (*iid*) random events. While this may be a suitable assumption in biology, it is certainly not so in the social sciences. Humans have the curious habit of combining conformity (i.e. of individually copying the relatively successful behaviour of others) with: (i) individual acts of subversion caused by some theory regarding the rules that govern their society (i.e. an ideology) and (ii) collective or coordinated acts of subversion intended clearly to undermine established social conventions and norms (e.g. confronting patriarchal notions of propriety, bourgeois norms of property rights). The conjunction of (i) and (ii) constitutes, in evolutionary

terms, behavioural patterns consistent with highly correlated mutations linked inextricably to the adaptation mechanism.

In short, (i) and (ii) disestablishes the joint assumption (A) and (B) without which the much-prized evolutionary economic models break down. Put differently, while humanity is typified by both *natural* and *social* selection, economics' evolutionary turn can only deal with the former. To the extent that human history is influenced systematically by our capacity for reflection, dialogue and political action (a capacity antithetical to the assumption of mutations as exclusively random *iid* events), evolutionary economics is insufficiently ... evolutionary.

53

To their credit, a number of evolutionary theorists have understood this well and tried to respond analytically.

54

However, they quickly reached the conclusion

55

that allowing the mutation probabilities to be cointegrated with the social adaptation mechanism yields a new type of *Folk Theorem*: i.e. almost *any* conventional behaviour can become disestablished and any alternative may take its place if 'subversives' coordinate their mutation probabilities appropriately and in response to the currently dominant behavioural conventions.

The wall of indeterminacy has, once again, defeated neoclassicism's efforts to rise to a new level of sophistication: Its attempt to infuse some realism into its models by borrowing heavily from evolutionary biology caused the set of (evolutionary) equilibria to divide and multiply *ad infinitum*.

56

In the face of such infectious indeterminacy, the mainstream recoiled, yet again, behind the strong version of its third meta-axiom (by insisting that mutations *are* random *iid* events

57

). This is unsurprising since its only other alternative would be to drop theoretical modelling and to concentrate either on simulations or on empirical work (or both). While some gallant evolutionary economists *did* focus on simulations (see Patokos, 2005), they soon realised that the mainstream left them behind, preferring to perform the **1→2→3→4** shuffle which took it back to a neoclassical position that is just as unsophisticated as the original (since the insistence that humans are incapable of coordinating their 'mutations' effectively returns us to a world of pseudo-rational fools).

Interestingly, in this case too, the theoretical failure enhanced greatly neoclassicism's discursive power courtesy of the new claim that its theorems can now be supported by an evolutionary narrative.

58

1.4 Behind neoclassicism's undiminished dominance

Neoclassicists are exceptionally open-minded people, willing to countenance *any* proposition, however farfetched, weird or even ... leftwing.

59

All they ask in return is that the said proposition is *embedded within their three meta-axioms*. This 'openness' is made all the more significant by the fact that, undoubtedly, *any* conceivable 'story' can be told by tinkering with neoclassicism's first two meta-axioms (see Dasgupta, 2002). Lured by the prospect of unbounded theoretical possibility, the aspiring young economist delights in tinkering her way into the infinite vistas of potential neoclassical narratives; she even revels in sailing the oceans of indeterminacy stirred up by her tinkering.

At some point, however, the fun must give way to publications, appointments and full induction into the profession. At *that* point, the lurking gatekeepers (supervisor, referees etc.) present her with a fresh condition: To be allowed into the priesthood, her models must have first achieved 'closure' (i.e. a restricted set of equilibria); she must, in effect, submit them to the merciless tightening of the third meta-axiom's fist, thus tracing the **r** or **b** trajectories (see the previous section's diagram) away from indeterminacy's *cul-de-sac*. At that juncture, having already invested great energy and hope in her modelling, it takes a brave and tragic theorist to desist and call it quits.

A tiny minority 'close' their models reluctantly, tucking critical comments away in their papers' footnotes, biding their time and, once tenured, turn into resident critics. Some 'close' their models and steer clear of any controversy, but nonetheless manage to retain the memory of how determinacy's imperatives whipped them back from a complex and rewarding inquiry to a paradigm devised for arid pure-exchange economies in which a sophisticated theory of agency, not to mention a left-of-centre political agenda, is as viable as a fire under a mighty waterfall (see Varoufakis, 2002, for the 'postmodern' aspect of this). Meanwhile, the vast majority not only leave no stone unturned to 'close' their models, often with moral enthusiasm, but also sweep under the emotional carpet any memory of how their models' 'closure' was bought at the price of returning *homo economicus* to strict isolation from his brethren, of relinquishing meaningful social norms, and of losing social and historical contingency.

Having performed the *dance's* moves *once* (with the **r** and/or **b** 'moves' back to positions **1** or **4**) in order to gain entry into the mainstream, the new recruits (the reluctant and the enthusiastic alike) soon discover that they must perform them again and again and again. For, once they are called upon to impart their wisdom in the amphitheatres, or to 'advise' government, business etc., their audiences *demand* a nuanced story of how their 'closed' models apply to the real world. Telling them that you can have *either* such a nuanced narrative *or* determinate models *but never both* requires the combination of intellectual honesty, mathematical acumen, and secure academic employment that only exceedingly rare birds, such as Nash or Debreu, possessed. In their absence, the vast majority sustain the illusion of a nuanced, determinate theory by keeping the *dance* going; by shifting backwards and forwards between 'closed' oversimplifications and complex-yet-indeterminate models; and, last but not least, by (sub-intentionally) hiding all this under a rhetorical cloak which gives (even to themselves) the impression of a serene, unchallenged scientific authority.

60

It is, of course, true that the very sight of a system of equations inspires a natural urge to solve it (and a feeling of disappointment when it proves over-determined). Non-neoclassicists (e.g. von Neumann, Sraffa, Goodwin, Robinson) are also subject to that

urge but, unlike the neoclassicists, did not *have* to sacrifice their theories' logical integrity in order to do so. Even the most mathematical amongst them (e.g. von Neumann), were relaxed with the idea of admitting exogenously determined variables into their analysis and introduced restrictive assumptions solely in order to solve their equations; not to 'close' their models shut.

61

Neoclassicists, in contrast, are hell-bent on the endogenous determination of *all* variables (prices, quantities, wages, profits; and even social norms, moral entitlements, psychological utilities) *exclusively on the basis of the initial, primitive data*. In short, they want to 'go it alone'; to reap the rewards of (social scientific) monopoly; to produce 'closed' theories packing historical, psychological, biological and anthropological relevance but with no input from meddling historians, uppity psychologists, boisterous biologists or doubting anthropologists. The three meta-axioms, in this sense, are enforced by the invisible hand of *academic rent seeking*; the same dynamic that motivates their *dance* as a device for maintaining the illusion of pluralist open-mindedness.

The question, however, remains: How does mainstream economics get away with this? Even if Kirman (1989) and Coase (1994) are right that professional economists have long stopped caring about the truth-status of their wares, does the world not notice their grand failure? I contend that it does. Students are abandoning economics majors in droves; the number of critical voices within the profession grows;

62

as for the public, official economic 'wisdom' causes derision or merriment. And yet, while academic economics is shrinking, the neoclassical stranglehold over the mainstream is as strong as ever. Why? I have already sketched out an explanation of what goes on *within* the discipline (our *dance of the meta-axioms*). But, there is a second reason relating to neoclassicism's immense ideological utility, viz. the current socio-economic order: Put simply, neoclassicism rules out *any* systemic analysis of capitalism.

Capitalism's champions have traditionally claimed that it is a *natural*, not a particular, *system*. Its critics (i.e. the Left) have objected that there is nothing natural about capitalism; that it is predicated upon a *particular* grid of political, legal and coercive power which could have been otherwise. Methodologically, this disagreement translates, simply, into whether really existing capitalism can be fruitfully theorised by models that keep structure separate from agency. Any economist who wants to breach the structure-agency separation

63

within neoclassicism's first two meta-axioms soon discovers that her models generate more equilibria than she can count. Thus, to continue a critical approach to capitalism she must either abandon the first two meta-axioms or accept indeterminacy. Either way, her papers will remain outside the mainstream.

In this sense, the profession's ostracism of *any* analysis that ventures beyond the three meta-axioms is tantamount to *a decree that every single mainstream economist accepts capitalism as a 'natural' system*.

64

Consequently, what we are left with is a profession churning out technical studies of fictitious markets which act as mere *diversions* from the real task of studying capitalism. Of course, the utility of this feat – for those who have an interest in keeping capitalism out of serious theoretical scrutiny – is immense. Capitalism appears in the public's eyes as a complex entity no less natural than the physical universe; it is, we are told, an entity to be analysed with the clinical impartiality of a social physicist,

65

exploited by financial engineers,

66

tamed by 'independent' Central Bankers, and only occasionally criticised by a few superannuated mainstream economists.

Recent neoclassicism and contemporary capitalism have given rise to a similar ontological claim: According to influential commentators, neither any longer exists! They are portrayed as gradually transcending into something altogether 'different'; of having, in fact, 'transformed' themselves out of existence.

67

Though this debate is well outside our paper's scope, it is tempting to note that the 'capitalism-has-disappeared' line of argument is jointly functional both to capitalism *and* to the dominance of neoclassical economics. It is functional to capitalism because it helps it remain invisible, shielding it from systematic criticism. And it is functional to neoclassicism because it justifies its insistence on the three meta-axioms.

While the world is currently struggling to make sense of the tumult visited upon it by a *particular* strand of globalising capitalism, the latter's best defence comes in the form of thousands of young economists being quick-marched headlong into academic obscurantism and socio-economic irrelevance. Instead of acting as the *avant-garde* that will prise out the truth about the causes and nature of the current crisis, they are conscripted to this perpetual feedback mechanism which mutually reinforces (a) the current economic order and (b) the neoclassical core of mainstream economics. Future historians, we suspect, will mark this out as our era's most fascinating, *and* most tragic, evolutionary social dynamic.

1.5 Epilogue

Neoclassical economics draws its immense narrative power from an audaciously circular process of mutual reinforcement: faithful to its constitutive meta-axioms, which it juggles continuously in a manner that hides their implications (and, often, their logical incoherence), neoclassicism retains its hold over the economics mainstream *and* rules itself out of engagement with the logic of really existing capitalism. The latter, supra-intentionally, rewards neoclassicism with institutional power which helps it maintain a strict embargo on any serious scrutiny of its own foundations.

It seems almost indelicate to point out that, while this feedback mechanism remains opaque and unexamined by the mainstream's critics, contemporary economic reality and mainstream economics will remain strangers who reinforce each other's dominance as long as (a) *mainstream economics remains, courtesy of its meta-axioms, innocent of the logic of capitalism* and (b) *the logic of contemporary capitalism spreads faster and deeper while economics' meta-axioms help it remain invisible*.

Quite possibly, never before has intellectual history fashioned an ideological triumph of this magnitude out of a sequence of sorry, yet powerfully motivated, theoretical failures.

1.5.1 A brief guide to the rest of the book

Turning to the rest of this book, every chapter that follows constitutes a case-study of – a personal experience with – what happens when ones attempts to 'civilise' a standard neoclassical model. At the beginning of each chapter a brief section links its theme with the analysis of neoclassicism's reproductive fitness that the present chapter just presented and introduces the reader to the class of theoretical models that it turns on. It also foreshadows the manner in which attempts to 'civilise' these models, to render them more realistic, resulted in radical indeterminacy. Then, at the end of the chapter, a 'chapter epilogue' classifies that particular case study in terms of the specific move (see

Section 1.3

above) of the *dance of the meta-axioms* by which neoclassicism ultimately sidestepped the indeterminacy and restored its authority, at the price of ensuring a complete incapacity to illuminate the real world phenomenon that that class of models

was meant to analyse.

Notes

- 1 For they think of what they do as scientific economics. The history of the term 'neoclassical' is discussed in Aspromourgos (1986). It should not be confused with the related term 'neoclassical synthesis' employed by Don Patinkin and Paul Samuelson to describe a reinterpretation of Keynes.
- 2 Victorian values and practices evolved through time and meant different things in different sub-periods; e.g. the late Byzantine era resembled its earlier more 'Roman' phase very little indeed. This dynamic complexity, however, does not detract from the usefulness of an over-arching characterisation such as 'Byzantine' or 'Victorian.'
- 3 A good example of such axiom-based definitions are Becker (1976), Blaug (1992), Vilks (1992), Hodgson (1999) and Colander (2005a). They define neoclassicism in terms of their *assumptions*. To take the most recent attempt to do so, Colander (2005a) defines neoclassicism, viz. the 'holy trinity' of rationality, greed and equilibrium. Notice that, in terms of his definition, all it takes for a theory to step outside neoclassicism is a minor relaxation of any of these axioms (a relaxation that every self respecting graduate student can perform in her spare time). It was, therefore, inevitable that Colander (2005b) would conclude that neoclassical economics is dissolving. In contrast, our meta-axiomatic definition accommodates evolving axioms which, while in flux, remain within what I think is a particular and highly distinctive method; one that not only 'survives' these relaxations, but in fact one that strengthens its stranglehold over the profession as it evolves. In this sense, our line of argument is more in tune with Dow (1995) and Fine (2008). But more on this in the next two sections.
- 4 Geanakoplos *et al.* (1989) offer an excellent case in point. By allowing an agent's utility to depend directly on her second-order beliefs regarding her own choice, as is the case more often than not for all of us (e.g. Jill's utility from passing an examination differs depending on whether she thought that Jack thought that she would pass or not), they enrich the model of individual agency. However, this enrichment comes at the price of indeterminacy even when the agent acts alone and under perfect information, viz. all relevant data (e.g. Jill's decision may belong to violently different equilibria; in one she studies hard expecting that Jack thinks he will pass, an expectation that she wants to fulfil; in another she thinks he is not expecting her to pass, a thought that makes her less eager to want to invest in this examination).
- 5 To mention a few, social norms have been allowed to 'infect' a worker's preferences in a manner that explains wage rigidity and even the decision to join a strike (see Akerlof, 1980; Varoufakis, 1989); preferences are formed endogenously (see Bowles, 1998); macroeconomic events influence individual motives (see Akerlof, 1982, 2007); social evolution determines private actions (see Weibull, 1995), what others think has a direct impact on what we want (see Rabin, 1993) etc.
- 6 Some non-neoclassical readers will protest that evolutionary game theory is not neoclassical. While I understand the hope this theory has given to many non-neoclassicists, and at the risk of wrecking it, I shall be arguing in the next section that evolutionary game theory remains firmly neoclassical (at least given the present section's definition of neoclassicism).
- 7 The strict separation of belief from preference relaxed, as in the case of psychological game theory – see Hargreaves-Heap and Varoufakis, 2004, Chapter 7. Weak methodological instrumentalism, see **s** below, accommodates such departures from **S**.
- 8 See Hargreaves-Heap and Varoufakis (2004), Chapter 6, for more.
- 9 Once upon a time, we could have instead talked of *methodological rationalism* as the dominant narrative centred on agents acting rationally. But since ordinal utilitarianism took over, there is no sense in narrating behaviour in terms of agents acting rationally. Instead, rationality is reduced to the consistency of one's preference ordering which, by definition, determines that which agents will do. See Arrow (1994) and Varoufakis (1998, Chapter 4).
- 10 See Varoufakis (2008) for the argument that such models are, essentially, ahistorical.
- 11 However, while **S**'s roots are Humean, Hume would have objected strongly to it. Our Reason, he would have thought, is too timid to tell us what is best in a social context, while our Passions are too unruly to fit neatly into some ordinal or expected utility function. It took the combined efforts of the late nineteenth century neoclassicists to build upon Jeremy Bentham's reduction of all the Passions to a single one (the passion for utility) before they tamed it sufficiently, bleached it of all psychology and sociality, thus reducing it to a unidimensional index of preference-ordering which is expressible as a smooth, double differentiable ordinal utility function.
- 12 In this sense, rather than being explained as the result of some complex calculus of the locals' desires, the logic of driving on the left in Gloucestershire, or on the right in South Maine, is to be found in some adaptation mechanism that followed on from a random event (or mutation), whose trace is often lost in the past, and which yielded a dominant evolutionary equilibrium.
- 13 'In every system of morality which I have hitherto met with ... I am surprised to find, that instead of the usual copulations of propositions, *is* and *is not*, I meet with no proposition that is not connected with an *ought* or an *ought not*. This

change is imperceptible; but is, however, of the last consequence.' Hume (1739/40, 1888; III, i, 1).

14 The obvious exception here is Keynes, who stands alone as a theorist committed to complete explanations of the workings of capitalism which are consistent with disequilibrium. See *Leijonhufvud* (1968).

15 While the neoclassicists' technical sophistication has taken off since the time of Cournot (and even of Arrow and Debreu), one truth remains: stability analysis is a fig leaf to cover up the dearth of any consistent theory of how a market equilibrium might emerge on the basis of historically situated acts of self interested buyers and sellers. In fact, as Mantel, 1974, and Sonnenschein, 1973, 1973, have famously shown, such a demonstration is impossible. Analogously, in game theory, the theorists' favourite equilibrium concept (subgame perfection) is also impossible to rationalise logically except under very special, atypical, circumstances (see Varoufakis, 1991, 1993).

16 Consider, for example, von Neumann's input-output analysis (von Neumann, 1937; a model that fits nicely in the classical economics tradition; see Kurz and Salvatori, 1993), the standard Sraffian model of determining prices in the context of joint production (Sraffa, 1975), Goodwin's dynamic equilibrium yielding a stable pattern of oscillating inflation and unemployment (Goodwin, 1967), Marxist schemas of reproduction (Halevi, 1998) etc. They all 'discover' the equilibrium state or path on the basis of their primitive data and some pre-specified selection mechanism (e.g. the assumption that profit rates will equalise across sectors).

17 For example, von Neumann's game theory (see von Neumann, 1928, and von Neumann and Morgenstern, 1944), while fully neoclassical, invariantly contained complete explanations of the reasoning that would lead players to equilibrium. Similarly with Marshall (1891), for whom equilibration was a process that required a comprehensive exegesis that is best attempted at a partial equilibrium level of abstraction.

18 For a complete account of how Nash's Cowles October 1950 presentation was the catalyst for Debreu's and Arrow's descent into formalism, and the ensuing static general equilibrium theory, see Varoufakis (2009).

19 General equilibrium theory's divorce from convergence analysis is well understood (see also note 15). Less appreciated is that a similar problem has been afflicting game theory ever since the Nash equilibrium became its foundational stone: While the simple, static Cournot-Nash oligopoly equilibrium requires no more than **e** to be arrived at, the moment the interaction acquires a more realistic structure (e.g. consists of a sequence of moves or is repeated) **e** does not suffice and **E** must be introduced urgently (and usually through the back door). See Hargreaves-Heap and Varoufakis, 2004, Chapters 2&3.

20 See Davis (2006), Colander (2005a, 2005b) and Colander *et al.* (2004a, 2004b).

21 Once upon a time, *Homo Economicus* was a simple lad (yes, a lad – see England 1993 and Hewitson, 1999). He liked what he bought and bought what he liked, loathed work, knew all he wanted to know (given the price of information), and cared not an iota either for his neighbours or for what they thought of him. As for the sort of economics built upon him, neoclassicism was typified by a familiar melange of theoretical practices: labour markets which would return to equilibrium if the troublesome unions and the meddling government let them; a habitual recourse to Say's Law; interest rates which never fail to equalise investment and savings; a constant array of Cobb-Douglas or CES production and utility functions; etc.

22 See Benrheim (1984) and Pearce (1984).

23 See Binmore (1989), Pettit and Sugden (1989) and Varoufakis (1993).

24 Take, for example, the standard prisoner's dilemma and suppose it is repeated indefinitely between the same players. The Folk Theorem shows that *anything* may happen as time goes by. Players may cooperate, they may defect, or they may oscillate between cooperation and defection in patterns of infinite complexity. By extension, this means that microeconomic theory has nothing to say regarding the formation or otherwise of cartels in oligopolistic markets: they may form, break down, reform at will and in ways that no neoclassical model can pin down analytically. See Hargreaves-Heap and Varoufakis (2004), Chapter 5.

25 The sheer convenience (for the modeller) of sticking to the assumption that rational agents must remain on the equilibrium path is aided and abetted by the fascinating, provocative, but ultimately deeply flawed, argument in Aumann (1976).

26 One such acknowledgment came from Levhari and Samuelson who in 1966 published a paper beginning with the admission that the neoclassical position was false: 'We wish to make it clear for the record that the nonreswitching theorem associated with us is definitely false. We are grateful to Dr. Pasinetti...' quoted in Burmeister (2000).

27 See Cohen and Harcourt (2003). See also Bliss (2005) for an illustration not only of the neoclassicists' readiness to ignore perfectly good scientific challenges but to take pleasure in taunting the challengers as well. He writes: 'If one asks the question: what new idea has come out of Anglo-Italian thinking in the past 20 years, one creates an embarrassing social situation. This is because it is not clear that anything new has come out of the old, bitter debates. Meanwhile mainstream theorizing has taken different directions. Interest has shifted from general equilibrium style (high-dimension) models to simple, mainly one-good models.' In one paragraph, Bliss depicts the challengers' incredulity that their perfectly valid challenge had no impact on the profession which recoiled

shamelessly behind the original, discredited neoclassical position.

28 Varoufakis (2009) argues that Nash's existence theorem in the context of games was the impetus which led Debreu and Arrow to their own proof of the existence of a vector of general equilibrium prices. This piece of 'speculation' was more recently confirmed by Kenneth Arrow himself who wrote: 'The [Nash] paper, however, supplied a firm basis by providing an existence theorem ...' (Arrow, 2009).

29 Debreu's background in the French Bourbaki mathematical tradition is consistent with a radical absence of any concern for the realism of his models (for an excellent account see Mirowski and Weintraub, 1994). Nash's bargaining theory can be seen as a precursor in this regard too in the sense of Nash's commitment to delivering a solution to the bargaining problem as long as he did not have to answer questions such as: 'How will they arrive at that bargain?'

30 See Bishop (1964) who tried to breathe a bargaining process, borrowed from Zeuthen (1930), into Nash's axiomatics. However, such attempts had the same basic flaw as that of Cournot's original, *circa* 1838, oligopoly dynamics: they assumed that agents would make assumptions which required a deep misconception of the model itself.

31 Debreu was always clear in his mind that out-of-equilibrium formalism is impossible. So much so that he, in fact, also rejected stability analysis: '(W)hen you are out of equilibrium, in economics you cannot assume that every commodity has a unique price because that is already an equilibrium determination.' (in Weintraub 2002). Nash, on the other hand, harboured hope that his formalism would be vindicated by some form of evolutionary analysis. In his PhD thesis he inserted a famous footnote in which he alluded to the idea of confirming his axiomatic derivation of equilibrium by positing players (drawn from a large population) who interact repeatedly (against a different opponent each time) without assuming that they '... have full knowledge of the total structure of the game, or the ability and inclination to go through any complex reasoning process'.

32 Assuming that delays in reaching agreement was costly to both bargainers.

33 For references see note 23.

34 In a nutshell, rational agents have no reason not to stray from 'the' equilibrium path (be it deterministic or stochastic) in a bid to subvert the expectations of their opponent for their own potential benefit. See Varoufakis (1991), Sugden (2000) and Chapter 6 of Hargreaves-Heap and Varoufakis (2004) for the complete argument.

35 A feeling that may be ameliorated better by defecting, rather than by cooperating.

36 She thinks that Jack expects her to cooperate. But since Jill knows that he knows, courtesy of CKR, that she is rational, she knows that he must also know that her decision to cooperate entails some sacrifice. Why would she sacrifice utility? The only explanation consistent with CKR is that she is *choosing* to forego some benefits in order to benefit him. Thus, if he responds by defecting, his choice reveals a degree of malevolence in the sense that it flies in the face of her 'kindness.'

37 Under (A) he is defecting on the common understanding that she will be doing likewise.

38 Suppose Jill predicts that Jack will cooperate. Under CKR, her only rational explanation is that Jack is prepared to sacrifice utility in order to benefit her. Her expectation that he is being kind to her puts her in a new type of dilemma: For if she defects, she will be profiting by trampling upon his kindness; a thought that may incur psychological costs for her. And if these costs are high enough, her best reply to his cooperative move is to cooperate too. On the occasion that both players hold similar beliefs, they may well find themselves in a new type of psychologically supported cooperative equilibrium which operates at three levels: actions, first-order beliefs *and* second-order beliefs.

39 Note that the direct reliance of players' utility function on second-order beliefs represents a switch to the weaker version of the first two meta-axioms.

40 All that is necessary in a standard static prisoner's dilemma to prove convergence to the mutual defection unique Nash equilibrium is the cast-iron logic of dominance reasoning: Whatever Jill (Jack) expects Jack (Jill) to do, she (he) is better off defecting.

QED This convergence mechanism falls within the ambit of version **e** of the third meta-axiom.

41 Note that the players' motivation (i.e. payoffs) can no longer be defined *a priori* as they depend on a combination of first- and second-order beliefs. Before Jill knows the utility value of mutual defection for her (in utility terms), she must know what to expect that Jack expects of her (and what she expects of him).

42 Mutual defection, mutual cooperation, Jill defects while Jack cooperates, and the latter's opposite.

43 I call it that because **E** must now impose equilibrium not only between acts and first-order beliefs but also between acts, first- *and* second-order beliefs. And it does this before the players get a chance to peruse the interaction! Thus the label **E** on steroids... *Methodological equilibration*, in this context, is no longer *prior* to *methodological individualism and instrumentalism* (as is the case in standard consumer theory, game theory or rational expectations macroeconomics); the axiomatic imposition of equilibrium is now necessary not just in order to predict the

interaction's outcome but also *in order to define the instrumentally rational agents' preferences!* (See Chapter 7 of Hargreaves-Heap and Varoufakis, 2004; Fehr and Gächter, 2000).

44 Notice how even this ultra-strong version of **E** has not defeated all the indeterminacy caused by the added psychological sophistication: In the end, the prisoner's dilemma, even after *a priori* assuming full alignment of actions, first- and second-order beliefs, now possesses two equilibria: One is the standard mutual defection outcome while the other is the cooperative outcome corresponding to mutually kind intentions (Jill expects Jack to cooperate in order to benefit her, thinking that she wants to do likewise; a thought which she is happy to confirm by cooperating herself).

45 The said leap is none other than the assumption that 1st and 2nd order beliefs are aligned *a priori*. It is, arguably, impossible to rationalise such an assumption as there is no logical explanation of how such alignment would ever come about (with commonly known certainty) in a static game.

46 It is less defensible because the version of the third meta-axiom it relies on stretches credulity beyond the limits of even the most impressionable neoclassicist. At the same time, it gains unprecedented discursive power due to the combination of: (a) the claims that neoclassicism no longer needs to posit psychologically unsophisticated agents, and (b) the immense complexity (which is necessary to model equilibrium behaviour in this type of analysis) which makes it *impossible* for anyone other than 'experts' even to understand the mathematical structure of the new type of model. The 'exclusion' of 'outsiders' lends power to the 'insiders' and evokes feelings of awe among the 'outsiders', including some who were hitherto critical of neoclassicism.

47 See Maynard Smith and Price (1974) and Dawkins (1976, 1980).

48 The vindication came from the demonstration that populations of mindless agents (who simply copy the more successful behaviour in their midst) converge onto equilibria that neo-classicists can only axiomatically impose on populations of hyper-rational agents. Nothing pleases the theorist more than the demonstration of a result's generality; especially when the same result is reached *via* wholly new paths.

49 I am referring here to the fact that the 'evolutionary turn' in fact produced greater accuracy by restricting the so-called 'equilibrium selection' problem. For example, it was demonstrated that evolutionary dynamics *always* lead to some Nash equilibrium but that, at the same time, not all Nash equilibria are consistent with evolutionary dynamics. In effect, the evolutionary turn has discarded some Nash equilibria, therefore restricting the 'equilibrium selection' problem and, in this manner, sharpening the theory's predictive accuracy.

50 More precisely, the **E** (strong) version of the third meta-axiom (i.e. simultaneously assuming CKR and common priors of belief) gave its place to weaker version **e** (i.e. a replicator dynamic 'copied' from Maynard Smith and Price, 1974).

51 Non-neoclassicists were seduced not only by the dropping of instrumental rationality and its extensions but primarily by the demonstration evolutionary adaptation mechanisms can yield hierarchies and discrimination on the basis of nothing more than *arbitrary* differences between agents. It took a small leap of the imagination to recognise this approach's potential for constructing a theory of institutionalised discrimination, even exploitation, within human society. See Hargreaves-Heap and Varoufakis (2002) for more on the joint evolution of conventions and discrimination.

52 Indeed, this is the view of, among others, Colander *et al.* (2004a, 2004b), Colander (2005a, 2005b), and Davis (2003, 2006).

53 One of the authors wishes to acknowledge useful discussions on this matter with Geoff Hodgson. He is, of course, not responsible for the resulting viewpoint.

54 To mention two relevant papers, Foster and Young (1990) acknowledge that politics is what happens when mutations are coordinated into aggregate shocks which test the established conventions while Kandori, Mailath and Rob (1993) examine the impact of rational experimentation in finite and discrete populations.

55 See Bergin and Lipman (1996).

56 For a fuller account see Hargreaves-Heap and Varoufakis, 2004, Chapter 6.

57 It did this in practice by focusing exclusively on evolutionary models where the mutation mechanism is utterly independent of the adaptation mechanism and agents are not allowed to attempt to pattern their mutations (either at the individual or the social level). This is equivalent to the Harsanyi-Aumann doctrine in game theory, to neglecting the SMD theorem in General Equilibrium, to turning to representative agent models in macroeconomics and so on. In short, it is another form of aggressively imposing version **E** of the third meta-axiom.

58 The discursive power emanating from claims to having established the evolutionary foundations of neoclassical equilibria would, of course, crumble under the weight of critiques like the one I presented above. However, neoclassicism is shielded from the force of such arguments due to their complexity. By elevating its failures at a higher level of abstraction, neoclassicism hides them from the eyes of all but a small minority who are keen (and able) to dwell into the hidden axioms. Sugden (2001) is one of that small minority. He coins the term 'slash-and-burn strategy' to

describe the manner in which economists approach non-neoclassical lines of inquiry, transplanting into economics ideas and concepts which were developed elsewhere, e.g. in biology, on the back of backbreaking empirical work. While proclaiming a profound interest in the work of biologists and others, in truth they have not a smidgeon of an interest in doing themselves any of the empirical work which would have been required to make the transplantation intellectually viable. For Sugden that is equivalent to slashing and burning a nearby forest by those who sing its praises.

59

See Elster (1982) and Roemer (1985, 1986) for some famous attempts to enlist neoclassicism to a leftwing cause.

60

McCloskey (1995) is the obvious source for insights into the mainstream's rhetorical strategies. Sugden (2001), in contrast, describes these practices more angrily: he calls it (recall note 58) a *slash-and-burn* strategy.

61

For example, the level of wages in Sraffa are exogenously varied, as they are in von Neumann's (1937, 1945) growth model. The latter, interestingly, was behind almost all facets of contemporary mathematical economics (from game theory to general equilibrium growth models to the use of fixed point theorems as tools for proving the existence of equilibria). Nevertheless, his economics is not, according to the definition in our paper, neoclassical (see Kurz and Salvatori, 1993; Mirowski, 2002; Varoufakis, 2009).

62

See Blaug (1992), Stiglitz (2002), and Fulbrook (2003, 2004) for a small sample.

63

For example, to allow for preferences not only to be endogenous but also contingent on expectations and social norms that are themselves comprised of higher order expectations and beliefs.

64

Consider, for example, the politically and philosophically charged notion of 'solidarity' and suppose one wants to examine it in a neoclassical light. In Arnsperger and Varoufakis (2003) we show that this is possible at the level of the individual (under the weak meta-axioms **d** and **s**) *as long as neoclassical 'closure'* (i.e. the **E** version of the third meta-axiom) *is not imposed*. The moment **E** is imposed, any meaningful conception of solidarity vanishes. Other examples are legion: Neoclassical sociology demonstrates the scope for neoclassical explanations of non-market 'social exchanges' within the family, the decision of a revolutionary group to refrain from blowing up a railroad bridge, the allocation of time to religious ceremonies within farming communities, and so on (see Becker, 1976; and Coleman, 1990). The formation of social institutions is modelled game theoretically with social norms sustaining gift exchange in traditional and modern industrial societies alike (see Akerlof, 1982; and Fehr and Gächter, 2000). However, all the interesting psychology, anthropology and sociology, built in these models upon the weak versions of the first two meta-axioms, is razed to the ground the moment we sneak in the strong version of the third meta-axiom for the purposes of yielding determinate equilibrium solutions. The latter are bought at the expense of assuming away all that is theoretically interesting, viz. the psychology of the persons involved and the nature of the social norms within their community. The feedback effects between preferences and norms, between predictions and motives, between actions and beliefs etc. are all sacrificed in pursuit of prediction. The special bond between parents and children, or revolutionaries, workers, NGO volunteers etc. is reduced to the type of bonds linking colluding oligopolists. In effect, such theories begin with great expectations, which they nourish in models relying on the first two meta-axioms, which are then set aside as we get down to the serious business of 'closing' the models by means of the third meta-axiom. The resulting theory is, thus, rendered methodologically consistent (within the ambit of the three meta-axioms of neoclassicism) by the same process that guarantees that they become (courtesy of the imposition of equilibrium conditions) well and truly *anthropologically inconsistent*.

65

Debreu, toeing a familiar neoclassical line (see Mirowski, 1989), declared himself proud that his Bourbakist mathematics liberated economics from ideology. In a recent interview he said: 'Moi, j'adopte simplement l'attitude suivante: que les hypothèses qui portent à des conclusions on peut en faire ce qu'on veut: si cela satisfait les économistes libéraux et les marxisants, parfait! Je ne peux rien demander de mieux. Intellectuellement vous êtes 25 *emporté par le courant des idées* et vous allez dans la direction où il vous porte.' (see Bini and Bruni, 1998). In Debreu (1986) he wrote: 'Foes of state intervention read in those two [welfare] theorems a mathematical demonstration of the unqualified superiority of market economies, while advocates of state intervention welcome the same theorems because the explicitness of their assumptions emphasizes discrepancies between the theoretic model and the economies that they observe.' However, what the above neglects is that, while the welfare theorems can, indeed, be interpreted differently by readers of different political persuasion, Debreu's method blinds all you adopt it to capitalism's particularities. And this is perhaps the greatest ideological interference *any* method could ever aspire to.

66

Where did the finance theorists behind the infamous credit default swaps (to mention one example) find the confidence to assume that default correlations would be low enough to stave off catastrophe? Varoufakis (2009, Section 4.3) argues that they found it in the same place where neoclassicists derived the confidence to impose the third meta-axiom (see

[subsection 4.3](#)

) every time they needed to 'close' one of their models.

67

Colander (2005b), for example, writes: '... previous views considered heterodox are moving into the mainstream, as the analytic and computing technology is allowing young researchers to develop these ideas in ways that will lead to institutional advancement... Because of these changes, today one would no longer describe modern economics as neoclassical economics.' (For more references along these lines see note 52.) Turning to capitalism, the respective line has for a while been that, due to technological change, the traditional analytical categories 'capital' and 'labour' have evolved to such an extent that it

no longer makes sense to define capitalism in the traditional manner.

2 Unity is strength

It is also the cause of indeterminacy regarding the wage and employment preferences of employers and trades unions

2.1 Prologue

2.1.1 Background briefing

My theoretical engagement with the theory of conflict began early on, as part of my PhD dissertation (Varoufakis, 1986) which involved delving into microeconomic tests of simple optimisation models of industrial action: of strikes, lock-outs, goslows etc. Those models (e.g. Ashenfelter and Johnson, 1969) assumed that at least one of the two sides of the employer-trades union negotiation was irrational (guess which!), resisting mechanically the other, the 'rational' party (whose decisions were, of course, instrumentally rational).

This asymmetry (one side rational the other myopic) allowed the modellers to cobble together an optimisation model according to which the 'rational negotiator' (invariably the firm) treated some exogenous resistance function provided by the 'irrational resistor' (the trades union) as its intertemporal constraint prior to solving for its optimal strategy on the strike duration v wage rise plane. In short, the trades union issued a wage rise demand at time $t = 0$ and then reduced it following some negative exponential during the strike. The firm, knowing this, simply selected the combination of strike duration and wage rises that maximised the present value of its profits, taking care to ensure that its short term strike costs (following resistance to the union demands) were nicely weighed up against the long term costs of settling the strike (i.e. the higher long term wage costs that a speedy resolution demanded).

At first sight, these models struck me as fascinating, mainly because of what they had left out. The first thing that I thought was missing was an acknowledgment that neither negotiating side owned a monopoly on rationality (or myopia, for that matter). As part of my dissertation I showed (this was the microeconomic part of my thesis) that the same data (concerning particular negotiations between corporations and trades unions) could be just as well explained regardless of whether we assumed (a) that the employer was the rational party and the trades union the myopic negotiator or (b) exactly the opposite (a rational trades union negotiating against a myopic employer).

Alas, recognising this brought along with it a hideous theoretical headache; the uncontrollable indeterminacy which emerged the moment Jill was allowed enough

rationality to calibrate her strategy on the basis of her expectations regarding what Jack thought that she expected him to believe that ... *ad infinitum*. In short, allowing both sides of a negotiation a degree of rationality – i.e. breaking down the existing literature's abrupt distinction between a clever bargainer confronted by a recalcitrant fool – meant the end of simple optimisation models and an inevitable foray into the realm of game theory. In the next chapter I narrate what happened when I decided to move in that direction after completing my PhD.

The second thing that was conspicuous by its absence from the literature was a provision for that which worries trades unions the most: imperfect mobilisation amongst workers. Indeed, the whole literature (e.g. Ashenfelter and Johnson, 1969; Farber, 1976; Akerlof, 1980) assumed that workers always acted as one person: either everyone joined a union and struck together when the need arose, or there was no union, no strike, nothing. At the time that I was researching this literature, the great tussle between trades unions, on the one hand, and the alliance of Mrs Thatcher's government with the employers, on the other, focused on the degree of unity the trades unions could muster. Suffice to remind the reader that, during the yearlong miners' strike, in 1984, all eyes were on the participation rate every single day: the proportion of

miners who did not cross picket lines in each shift, a datum that was being monitored by the combatants and reported by the media with abandon. It was when that rate dropped below a certain level that the miners' defeat was declared. So, I decided to get down to work to see if the existing models could be made more realistic by allowing for different, endogenous, levels of worker mobilisation. The result is the kind of analysis in the present chapter.

The link with the book's theme is intimate. This chapter puts on display what happened when, as a green-behind-the-ears researcher, I enthusiastically attempted to relax an assumption that prevented the existing, mainstream literature from telling a reasonably realistic story: the assumption of automatic, perfect worker mobilisation whenever their trades union calls a strike. At the outset, I embarked with the optimism of all great beginnings. I thought that I could demonstrate the analytical value of introducing worker unity (or disunity) as an important endogenous variable to models of industrial disputes. Little did I know that the spectre of indeterminacy would always be lurking, ensuring that the economics profession would, in the final analysis, reliably perform the *dance of the metaaxioms* in a manner that thwarts all attempts to civilise our models.

2.1.2 Chapter guide

Section 2.2

, based on Varoufakis (1989) presents the simplest model that imbues existing one-sided optimisation models with the notion of imperfect worker mobilisation or solidarity. In it, I assume that employment is fixed and the only issue under negotiation is the wage rate. The model's main contribution is to model worker solidarity or mobilisation as an endogenous variable that reflects a social convention, within the community of workers, which determines the psychological costs to each worker from breaking the strike (i.e. from crossing pickets lines). By the end, it becomes clear that the success of a trades union's campaign, and thus

the final wage outcome, cannot be understood on the basis of 'hard economic data' alone. Indeed, a major determinant of the wage outcome is the dynamic feedback between workers' moral beliefs about crossing picket lines and the trades union's bargaining power (which, naturally, is a function of the mobilisation rate).

Section 2.3

generalises this model by allowing for the obvious: for trades unions and employers to negotiate both on wage and on employment levels. When this happens, in the context of endogenous, imperfect worker solidarity/mobilisation, the spectre of indeterminacy spreads. Suddenly not only is it impossible to predict the negotiation's outcomes but, remarkably, the trades unions' own targets (regarding their preferred wage and employment levels) are also indeterminate – see Varoufakis (1990a). In short, mainstream models of wages and employment determination implode; a message that, as you may imagine, the 'profession' was unprepared to welcome.

2.2 Worker solidarity and strikes

1

2.2.1 Introduction

Strikes usually succeed or fail depending on the degree of solidarity the trades union can muster. And yet the strike literature continues to assume perfect participation by the workforce and to treat capital-labour bargaining as a two-person division game.

2

In contrast to other areas of economic analysis, where egoistic behaviour by agents may inhibit the provision of public goods, the theory of industrial conflict has failed to articulate a sound explanation of why individuals may choose to cooperate in the pursuit of common objectives. Moreover, when wage determination is consistently examined as if trades union unity is never an issue, one wonders what aspects of wage setting

remain hidden.

In previous work, Varoufakis (1986) and Naylor (1987) adapted Akerlof's (1980) social custom model to shed some light on how a strike coalition may come into being. The common theme in these models has been the willingness to visualise the worker as a social animal lacking access to a genuine free ride. By assuming that those who cross picket lines incur socially, as well as privately, determined reputation-*cum*-psychological costs, defection loses much of the appeal it would have had in a world inhabited by individuals who live in splendid isolation from one another and, more generally, from their social environment. Under certain circumstances, it is shown that rational agents will choose to sacrifice monetary gains in exchange for valuable reputation and/or the psychological inner glow from having acted in solidarity with others. The trades union's public good, i.e. the strike, will therefore be, at least partially, provided.

In what follows I shall be building on these earlier results to endogenise fully wage and strike duration outcomes.

Section 2.2.2

builds a version of the original social custom model and discusses the existence and nature of solidarity equilibria. The stability of worker coalitions is examined and it is shown that, in the face of multiple equilibria, the outcome depends on initial beliefs as well as on the heterogeneity of preferences. In

Section 2.2.3

I introduce the two negotiating teams: trades union leaders who are fully informed about the level of the firm's product

demand, and managers who are imperfectly informed on the trades union's internal cohesion. In order to provide the firm with an incentive to divulge privileged information, the trades union's leaders offer wage/strike duration combinations from which the firm has to choose one. If the chosen bundle involves a positive strike duration then conflict occurs (i.e. a strike ensues) and its purpose is to convey optimally information to the trades union.

Section 2.2.4

to

2.2.6

blends the models of

Sections 2.2.2

and

2.2.3

and examines how variable worker solidarity can give rise to dynamic adjustments in employer offers. In

Section 2.2.7

I demonstrate that an interesting three phase history of the trades union-firm relationship emerges once bounds are placed over the bargaining horizon. Unsurprisingly, the model's main characteristic turns out to be the sensitivity of outcomes to the properties and timing of the strike coalition.

Section 2.2.8

sums up the findings.

2.2.2 There is no such thing as a free ride: the emergence of a strike coalition

As trades union gains are reaped by both workers who make short term sacrifices during a strike and those who free-ride, collective impotence (rather than the assumption of perfect mobilisation) should be taken as the starting point. One way of explaining why free riding is frequently defeated is to argue that free-riding behaviour ('defection' in game theoretical parlance) carries its own legacy and ceases to be a ride devoid of cost. Provided a social ethic is in place which prompts society to view

defection as deplorable, all that is needed for rational cooperation to emerge is the assumption that individuals suffer disutility from being identified as disloyal to the common cause. In that case, workers may, in customary neoclassical fashion, choose to strike if doing so enhances their utility. The following assumptions set the scene:

ASSUMPTION 2.1 *An individual worker decides on whether a strike is worth joining as part of an exercise in personal utility maximisation*

ASSUMPTION 2.2 *Workers draw utility from two sources: expected income and from the psychological rewards from interacting with their colleagues, including their reputation amongst their peers*

ASSUMPTION 2.3 *The degree of disutility associated with strike-breaking depends on both the utility from expected income and from the psychological utility of breaking, or joining, the strike.*

The first two assumptions can be expressed in terms of the typical worker's utility function $U(w, \Psi)$ with w the wage rate and Ψ the psychological rewards at any point in time, the latter being determined thus:

$$\Psi = \begin{cases} \Psi_j & \text{when the worker joins the strike} \\ \Psi_{nj} & \text{otherwise} \end{cases} \quad (\text{with } \Psi_j \geq \Psi_{nj})$$

Letting

T be the length of the currently negotiated contract

s the expected strike duration

w the expected wage settlement

r the discount factor of all agents in the model

b the level of strike benefits per period (payable by the trades union, the community etc.), and

w_0^e the employer's final pre-strike offer ($w_0^e > b$)

our worker will join the strike beginning at time 0 and expected to last until time s iff

$$\int_s^T U(w, \Psi_j) e^{-rt} dt + \int_0^s U(b, \Psi_j) e^{-rt} dt > \int_s^T U(w, \Psi_{nj}) e^{-rt} dt + \int_0^s U(w_0^e, \Psi_j) e^{-rt} dt$$

The gist here is that there are pros and cons to whether the worker joins or not (thus, the assertion in this section's title that 'there is no such thing as a free ride'). The first integral of both the l.h.s. and the r.h.s. of the inequality above relate the worker's long term utility, once the strike has ended. The only difference between these two integrals (the first of each side) is due to psychological effects, as the wage enjoyed by strikers and strike-breakers, once the dispute is over, is the same. In short, strike-breakers carry into the future whatever 'psychological losses' have been incurred from having crossed picket lines (even if they are discounted at rate r). Turning now to the second integral of each side, this captures the worker's utility *during the strike*. Here, both income and psychological effects differ. Strikers 'enjoy' a psychological (or 'moral') utility advantage ($\Psi_j \geq \Psi_{nj}$) but suffer an income loss ($w_0^e > b$).

The above decision rule is made more manageable, without loss of generality, by letting $T \rightarrow \infty$ and adopting a simple separable utility function of the form $U(w, \Psi) = \alpha_1 w + \alpha_2 \Psi$. It is now simple to show that the above inequality, i.e. the worker's decision rule, reduces nicely to:

$$\begin{array}{lll} \text{Decision Rule} & \begin{array}{l} \text{join} \\ \text{strike break} \\ \text{indifferent} \end{array} & \begin{array}{l} \text{if } \rho\gamma > z \\ \text{if } \rho\gamma < z \\ \text{if } \rho\gamma = z \end{array} \end{array} \quad (2.1)$$

where $\gamma = \alpha_2/\alpha_1$ is the worker's relative valuation of psychological rewards, $\rho = \Psi_j - \Psi_{nj}$ the magnitude of reputation loss from not joining and

$$z = (w_0^e - b)(1 - e^{-rs}) \quad (2.2)$$

The simple rule in (

2.1

) confirms the intuition that the probability of participation is positively related to the relative weight one attaches to the psychological rewards from joining the strike and the trades union's ability to compensate

its members during the dispute, while inversely related to the firm's final offer and the expected duration of the strike. Less obvious predictions can now be generated once we recognise the importance of group dynamics. In particular, I shall assign an important role to the expected degree of mobilisation within the trades union; i.e. the proportion of workers willing to join the strike, say, σ , $0 \leq \sigma \leq 1$.

Function z can be thought of as the rule's component encapsulating the decision's monetary considerations and will be therefore referred to as the 'monetary' part. It is clear that the average worker expects a shorter strike to be sufficient for the attainment of a given wage demand the greater the degree of mobilisation by the trades union. Moreover, with more and more workers ready to join in, the employer will increase the pre-strike offer, w_0^e , as a further inducement to strikebreaking. Thus, both w_0^e and s are treated by workers as a function of the level of solidarity or mobilisation σ . In particular, the rate of change in z with respect to changes in solidarity is given by

$$z' \equiv \partial z / \partial \sigma = \frac{\partial w_0^e}{\partial \sigma} (1 - \exp(-rs(\sigma))) + [w_0^e(\sigma) - b] r (\exp(-rs(\sigma))) \frac{\partial s(\sigma)}{\partial \sigma} \quad (2.3)$$

The greater (a) the employer's final offer and (b) the responsiveness of the expected strike duration to anticipated levels of solidarity, the flatter the z - σ relation in

Figure 2.1

below. The significance of this will become apparent when formulating

Proposition 2.1

Having codified

Assumptions 2.1

and

2.2

in rule (

2.1

), we can now incorporate

Assumption 2.3

. This is straightforward in view of the logical dependence of the psychological, or subjective, part of (

2.1

) – i.e. $\gamma\rho$ – to solidarity. Indeed, there is every reason to believe that a custom respected by very few is a weak custom

with limited capacity to influence the individual. Following Akerlof (1980), I shall therefore postulate that the larger the coalition of strikers the greater the stigma attached to crossing picket lines. In short,

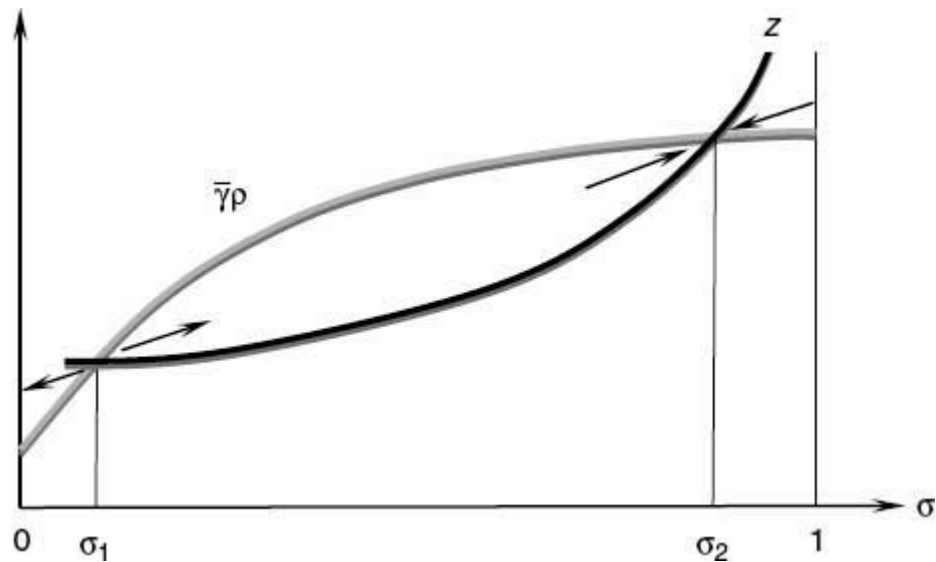


Figure 2.1

Two equilibria: perfect mobilisation or none at all.

$$\rho = \rho(\sigma), \rho' > 0 \quad (2.4)$$

Equilibrium solidarity (definition) – A coalition of workers is said to be in equilibrium if neither those workers who have decided to join it nor those who have chosen to continue working have an incentive to change their minds.

In the case of identical workers, the condition for a coalition $\hat{\sigma}$ to be in equilibrium is that $z(\hat{\sigma}) = \gamma \rho(\hat{\sigma})$ with intra-marginal strikers and strike-breakers alike being indifferent between changing sides and staying put. The same result applies when personal tastes differ between workers but are uniformly distributed over a continuous range. However, if there exist different sub-groups within the trades union holding diverse ideological positions, with γ values to match, an equilibrium may arise which involves no marginal workers. For a brief discussion see the Appendix to this chapter.

PROPOSITION 2.1 A sufficient condition for an interior equilibrium $\hat{\sigma}$ (i.e. $0 \leq \hat{\sigma} \leq 1$) to be stable is that $z'(\hat{\sigma}) = \gamma \rho'(\hat{\sigma})$.

From

Figure 2.1

it is clear that, when z cuts $\gamma \rho$ from below, any perturbation around the interior equilibrium σ_2 will trigger rule (

2.1

)
3

and return us to it. This is, however, not the case with σ_1 .

The inequality of slopes crucial to

Proposition 2.1

would have been a necessary, in addition to sufficient, condition for stability if workers were either identical or their differences were uniformly distributed. However, more unevenly distributed personal valuations of the psychological aspects of this decision could generate stable equilibria even if the condition in

Proposition 2.1

is violated. Regarding the distribution of the γ 's, I proceed with a simplifying assumption:

ASSUMPTION 2.4 $\gamma_i \sim I(\gamma_0, \gamma_1) \quad \forall i = 1, \dots, n, \gamma_0 < \gamma_1, \quad \bar{\gamma} = \frac{\gamma_0 + \gamma_1}{2}$,
where n is the size of the trades union's membership and γ_0, γ_1 correspond to the valuations of the psychological effects of striking, or not, by the least socially motivated

and by the most loyal union member respectively.

PROPOSITION 2.2 *The trades union's preparation for the strike will result in an imperfect level of solidarity only if the employer is expected to 'take' a strike irrespective of the union's mobilisation rates.*

The above asserts that when perfect solidarity is expected to win the day for the trades union, there will be no dichotomy between its members: collective rationality will prevail as the workers either stand solidly behind their trades union or unanimously refuse to down tools. Free-riding will, therefore, only become a possibility if it is known that absolute unity is not sufficient to force the employer into

submission. Therefore, the contradiction between private and collective interest does not manifest itself unless sacrifices are called for.

Proof: The condition for a stable equilibrium at $\hat{\sigma}$ can be written as:

$$\frac{\partial w_0^e(\hat{\sigma})}{\partial \sigma} (1 - \exp(-rs(\hat{\sigma}))) + [w_0^e(\hat{\sigma}) - b]r(\exp(-rs(\hat{\sigma}))) \frac{\partial s(\hat{\sigma})}{\partial \sigma} > \bar{\gamma} \frac{\partial \varrho(\hat{\sigma})}{\partial \sigma} \quad (2.5)$$

If expected strike duration vanishes as σ increases (i.e. $s \rightarrow 0$ when $\sigma \rightarrow 1$, or $s = 0$ when σ hits an upper bound), inequality (

2.5

) reduces to $r[w_0^e(\hat{\sigma}) - b] \frac{\partial s(\hat{\sigma})}{\partial \sigma} > \bar{\gamma} \frac{\partial \varrho(\hat{\sigma})}{\partial \sigma}$, a condition that cannot be fulfilled since $\rho' > 0$, $s' < 0$ and $w_0^e(\hat{\sigma}) > b$. If, on the other hand, a positive strike length is thought necessary even when $\sigma = 1$, the mixture of strikers and free-riders will be stable provided:

$$\frac{\partial w_0^e(\hat{\sigma})/\partial \sigma}{\partial \varrho(\hat{\sigma})/\partial \sigma} > \frac{\bar{\gamma}}{1 - \exp(-rs(\hat{\sigma}))} \quad (2.6)$$

The multiplicity of outcomes calls for an interpretation of observed solidarity, or mobilisation, levels. A useful way of thinking about them is as equilibria of beliefs. Suppose that in the context of

Figure 2.1

workers initially expect solidarity to be below σ_1 . In that case, the model's dynamics will ensure that no worker will want to intend to strike. Conversely, if the typical expectation exceeds σ_1 , equilibrium behaviour commends σ_2 as the dominant coalition size.

2.2.3 The strike

The moment workers walk out, the preceding analysis is altered in two important ways.

First, $w_0^e(\hat{\sigma})$ is disengaged from σ as any further concessions by the employer are suspended until a final agreement is reached, or the trades union caves in. Secondly, workers may alter their relative valuations of the psychological effects of their behaviour during the strike (γ) when they gain first hand experience of the hardship involved, but also of the cathartic properties of collective action. The effect of the former is to cause instantaneously a rotation of the z function – see

Figure 2.2

– as $\partial w_0^e(\hat{\sigma})$ is set to zero and thus leading to a potential destabilisation of the coalition [if $z'(\hat{\sigma}) < \bar{\gamma} \rho'(\hat{\sigma})$]. In that case, the effect of the strike on $\bar{\gamma}$ becomes central to the fortunes of the strike. If the new $\bar{\gamma}$, say γ' , exceeds the old, then $\gamma \rho$ moves upwards and a bandwagon effect takes σ to unity. In the opposite case, when the strike adversely affects the psychological subjective benefits from participating in it, solidarity will wither. A possible model for this dynamic adjustment is given by

$$\dot{\sigma} = \lambda[\rho(\sigma) - z(\sigma)] \quad (2.7)$$

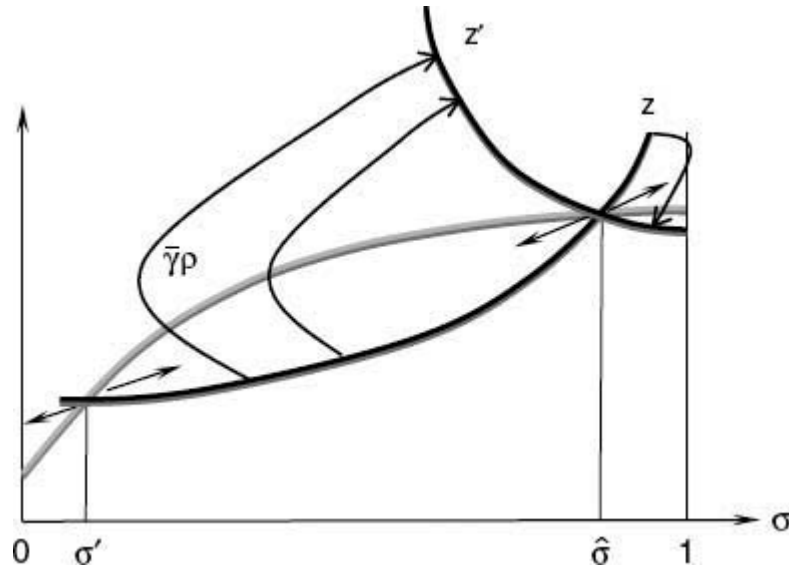


Figure 2.2

The effect of strike action on the actual willingness to mobilise. At time $s = 0$ a proportion of $\hat{\sigma}$ strikers walk out and the z -function rotates to its new position z' , rendering $\hat{\sigma}$ an unstable equilibrium. At that point, depending on the sign of the perturbation, either all workers will walk out or the strike will collapse.

2.2.4 A model of strike causation

In mainstream economics strikes are perceived as either a manifestation of irrationality (see Cross, 1969; Ashenfelter and Johnson, 1969) or the unavoidable cost of informational asymmetries. In what follows I shall adopt the latter interpretation, even though (I shall be arguing in the following chapter) it can be shown to be problematic. For now, let us keep faith with the standard analysis according to which conflict is viewed as an information transmission mechanism by which one bargaining side signals to another its true intentions. Trades unionists are assumed less well informed than managers regarding the state of demand for the final products. Since it is always in the interests of employers to pretend that the market is depressed, their readiness to suffer the costs of industrial action is perhaps the only way that the trades union can be convinced that demand is, truly, low. In short, a firm may have no other means of conveying to the trades union information that the latter lacks than by 'taking' a strike.

Hayes (1984) showed that when the trades union is seeking a wage demand that maximises its expected utility subject to the firm's optimal response, a perfect equilibrium union strategy is to present management with a schedule of wage demands that is a decreasing function of strike duration and independent of the true state of product demand. It is worth noting that the crux of Hayes' model revolves around the credibility of this schedule. Unless the trades union can articulate a believable pledge that it will obstruct renegotiation of the initially presented locus of proposed wage outcomes (on the wage-strike duration plane), the model

collapses. Although such a degree of commitment appears to be suspect, there is an impeccable logic behind it provided the trades union and the firm expect to bargain periodically, and indefinitely, under similar circumstances. The effect of relaxing the stringent commitment requirements are discussed in

Section 2.2.5

below.

During pre-strike negotiations, the trades union presents the firm with a concession schedule (CS) representing the collection of outcomes out of which management must select a single one:

$$\text{CS: } w^u = f(\hat{V}, \xi \hat{\lambda}, s), \text{ with } f'_V > 0, f'_\xi > 0, f'_s < 0 \quad (2.8)^4$$

where

$$\hat{V} = \begin{cases} 1 & \text{if } \gamma' > \bar{\gamma} \\ -1 & \text{if } \gamma' < \bar{\gamma} \end{cases}$$

w_u is the trades union's wage demand

its leaders' estimate of the firm's total revenues, while

reflects the unionists' views on the direction of the bandwagon effect once the pre-strike equilibrium is destabilised in

Figure 2.2

(i.e. on whether solidarity will reach 100 per cent or dwindle inexorably toward zero)

Having no other alternative, employers treat (

2.8

) as the constraint subject to which they must now optimise their firm's intertemporal profits. In short, the firm selects its preferred bundle of wages (w) and strike duration, from the CS function in (

2.8

), which maximises (

2.9

) below:

$$\int_s^T [V(L) - wL]e^{-rt} dt + \int_0^s \{V[(1-\sigma)L] - w_0^e(1-\sigma)L\} e^{-rt} dt - \int_0^T F e^{-rt} dt \quad (2.9)$$

where

$V[\cdot]$ is the firm's total revenue

F are the per period fixed costs

L is the (presumed) fixed level of employment in the firm

The firm's optimisation will favour a strike iff V is below a certain threshold level. This level corresponds to the one that, the moment firm finds itself below it, it is in its long term interests to 'take' a strike rather than to settle [by granting the wage level in (

2.8

) corresponding to a zero strike duration].

5

Simplifying the firm's profit function during the strike as $\int_0^s m(\sigma)[V - w_0^e L]e^{-rt} dt$, where $0 < m(\sigma) < 1$ is the average proportion of normal profits generated during the stoppage by strike-breakers and evaluated as wage rate w_0^e , letting $T \rightarrow \infty$, and imposing the exponentially declining function in (

2.10

below) for the CS in (

2.8

) above,

$$w^u = w_* + (w_0^u - w_*) \exp(-as) \quad (2.10)$$

(where w_* is the previous wage rate, w_0^u is the initial trades union wage demand that will avert the strike, and a is speed with which the trades union reduces its wage demand once the strike has commenced) yields the employer's optimal wage target:

$$w = \frac{a}{a+r} w_* + \frac{r}{a+r} \left\{ \frac{V[1-m(\sigma)]}{L} - m(\sigma) w_0^e \right\} \quad (2.11)$$

which will be agreed upon after a strike lasting

$$s = \frac{\ln(w_0^e - w_*) - \ln(w - w_*)}{a} \quad (2.12)$$

Not surprisingly, the final wage will be higher the greater the actual level of the producer surplus, the better the reputation of the trades union as a cohesive organisation, and the higher the previous wage rate.

Equations (2.11)

and (

2.12

) depict the outcome as perceived by the employer side at $s = 0$. If (

2.12

) reports a non-negative value, a corner solution averts the strike. Alternatively, the dispute goes ahead with σL workers walking off the job. It is the evolution of this coalition that plays the major role in determining any deviations of the final outcomes from those predicted by the static model.

2.2.5 Mobilisation, solidarity and strike dynamics

Unlike the trades union, employers can adjust their offers along the CS to suit new information on the level of solidarity. Any deviation of $m(\sigma)$ from the anticipated level will overturn the optimality of (

2.11

) and (

2.12

) and prompt managers to reconsider. There are three obvious parameters whose re-estimation may lead to revised offers.

First, right at the strike's outset (at $s = 0$), the employer will be keen to observe whether the initial level of mobilisation lives up to the *ex ante* edictions. Such alterations will lead to one-off discrete changes in the firm's offer.

Secondly, in view of the likely destabilisation of equilibrium solidarity (recall the rotation in

Figure 2.2

at the moment the strike begins), employers will be wise to keep a wary eye on the direction of the bandwagon effect (i.e. on ξ). If $\xi = 1$, the firm's worst fears will have been realised as the strike gathers momentum. Suppose that, for instance, the firm's *ex ante* expectation of ξ is $\xi_e = -1$ while, once the strike got underway, $\xi = 1$. In this case, the firm has got its predictions badly wrong with mobilisation and solidarity amongst workers strengthening monotonically rather than withering away as expected by managers. The firm will soon scramble to produce a revised, more generous, wage offer.

Thirdly, even if $\xi_e = \xi = -1$, there is still the matter of the speed (λ) with which worker mobilisation ebbs. To illustrate, suppose that the employer side held over-optimistic expectations about λ . Coming to terms with the trades union's higherthan hopedfor ability to impose strike costs will necessitate a shift in the firm's isoprofit firms on the (w, s) plane, the effect being a higher wage offer and

a shorter strike. Let $\pi(\lambda)$ be the firm's prior of belief concerning the value of λ . When observed mobilisation rates are at level σ , and the conditional distribution of σ is $f(\sigma | \lambda)$, the posterior distribution of beliefs is derived by Bayes Rule as

$\pi(\lambda | \sigma) = \frac{f(\sigma | \lambda) \pi(\lambda)}{p(\sigma)}$, where $p(\sigma)$ is the Bayesian prior distribution of beliefs concerning σ . It can now be shown

that the Bayesian updating mechanism or the firm's point estimate of λ after s periods of the strike is given by (

2.13

) below:

$$\hat{\lambda} = \iint \lambda f(\sigma | \lambda) \pi(\lambda) d\sigma d\lambda \quad (2.13)$$

Consequently, when $\hat{\lambda}$ is exposed as an over(under)estimate, managers will, provided $\xi = -1$, increase(decrease) their offer accordingly.

In geometric terms, the firm's isoprofit map, as defined on the (w, s) plane, rotates and its new tangency points with CS constitute the new intertemporally optimal strategies for the firm.

Figure 2.3

depicts one possible scenario.

2.2.6 Finite horizons plus a little uncertainty spawn three phases

The sceptical reader will rightly remark that the whole model collapses the moment the credibility of the CS is undermined. If there were no room for relaxing this requirement, then the generality of the analysis would be in jeopardy. The most obvious way of establishing the truth behind this suspicion is by removing the justification for the unwavering trades union commitment to the CS, namely the infinite horizon.

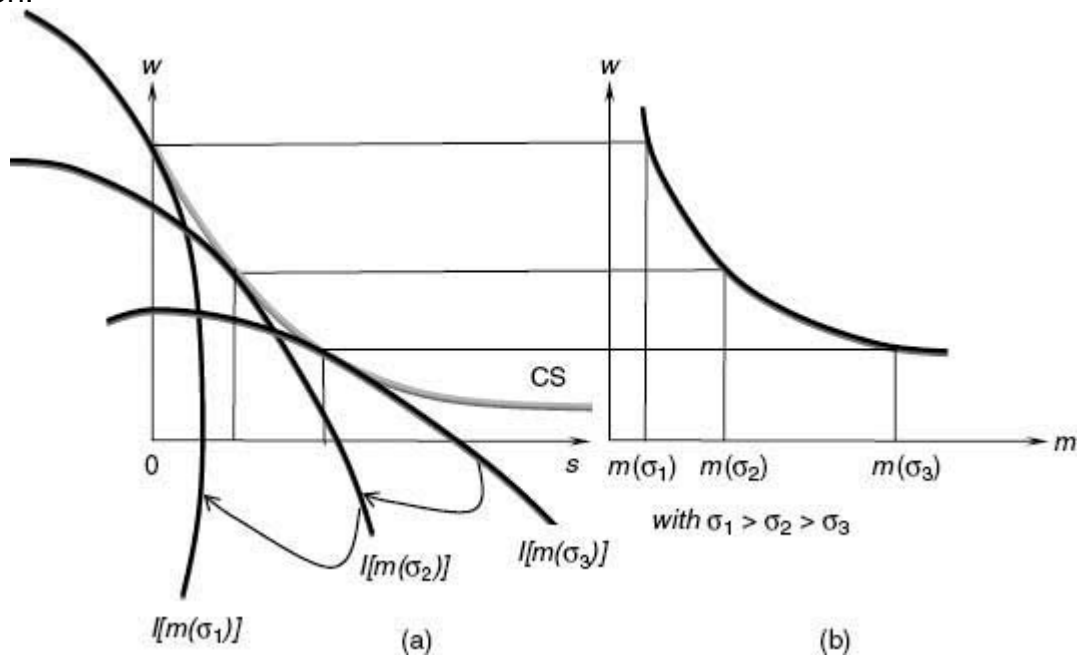


Figure 2.3

How expected mobilisation determines the employers' wage expectations. The l.h.s. diagram comprises the CS from (

2.8

) plus the firm's highest isoprofit curves, one for each level of solidarity/mobilisation σ . The r.h.s. diagram illustrates the projected wage outcome in relation to average mobilisation. As mobilisation estimates rise, from a lowly σ_3 to σ_2 and then to σ_1 , the firm's isoprofit curves corresponding to maximum inter-temporal profits rotate in the l.h.s. diagram, thus yielding higher wage offers.

Recall that the trades union is assumed to adhere to the CS even if during the dispute its solidarity predictions were found wanting. In the short run, union leaders would wish they were able to alter the CS, since the beliefs on which it was predicated proved erroneous. However, doing so would allow the firm to toy with the idea of ignoring the trades union's declared CS in future negotiations. This would damage

irreparably the labour force's long term bargaining power. Thus, an infinite horizon fosters absolute commitment to the CS.

What if, however, an upper bound, be it fixed or stochastic, were to be placed over the horizon? For instance, the current trades union leader may be due to retire after, say, k negotiations? Will the analysis require major revisions? Consider the k th negotiation. Leaders who are solely interested in pursuing the trades union's objectives will have no hesitation in rethinking the CS during the strike. Doing so will have no long-term effect since the opposite side knows that the current negotiation is the last they will conduct prior to retirement. Hence, the model of preceding sections is made redundant by the mere fact that employers are not convinced that during the k th negotiation the trades unionists will stick to their CS once the strike has commenced. Furthermore, if this breakdown is expected to occur in the k th negotiation it will also be anticipated in the $(k - 1)$ negotiation. Backward induction unravels and it quickly transpires that at no time will our onesided asymmetrical information model represent a fair account of the bargaining process!

Thankfully, all is not lost. The bulk of the analysis can be revived by injecting a small amount of uncertainty in the mind of the employer concerning the trades unionists' preferences.

8

This is how: backward induction unfolds only because its logic achieves a toehold during negotiation k . It is the conviction that the trades union will certainly wish to drift away from its own CS (once the strike has began) that leads to the model's collapse sequentially in all prior negotiations. In an environment where the unionists' motives are transparent, the trades union loses its ability to bargain credibly. However, if there is some doubt about the union's behaviour at time k , then this is not so. Suppose that, indeed, there is a small probability that the leader is averse to going back on a pledge (i.e. the CS) even if doing so were to increase the trades union's payoffs. Let's call this type of unionist 'intransigent'. The slightest of doubts as to whether the union's leadership is intransigent suffices to throw a spanner in backward induction's works and gives rise to a sequential equilibrium as follows:

PROPOSITION 2.3 *After the election of the trades union's leader to a fixed term in office, three distinct phases of stochastic length arise:*

Phase 1: Bargainers behave as if the horizon is indefinite. With the firm reluctant to contest the trades union's CS, the analysis of the previous sections applies intact.

Phase 2: The transition to this phase occurs when the time comes for the firm optimally to challenge the trades union's commitment to its own CS. As long as the trades union remains defiant in its commitment to its CS, wage and strike duration outcomes will continue to fall on the CS (see the l.h.s. of

Figure 2.3

). Thus, the original analysis only requires minor adjustment to allow for the longer strike durations, and lower wage awards, necessary to rebuff the employer's challenge.

Phase 3: Provided labour leaders would prefer to concede during the strike if it were not for the long-term repercussions of such acquiescence, they will eventually do so as their term in office draws to an end. The period of credible commitment to ex ante demands is now over and a qualitatively different model of strikes is needed to guide us through the uncharted third phase that leads to fresh trades union elections.

The employer's dilemma, brought on by the introduction of the finite horizon, revolves around the wisdom of pressing for a settlement off the CS. If it challenges the CS by insisting on such a bundle, e.g. on the combination of (w', s) in

Figure 2.4

, there is a risk that the trades union will wish to teach it a lesson by only accepting that particular wage after a longer dispute, i.e. aiming for combination (w', s') , as a means of shoring up its commitment to its own CS. Of course, this risk is only a problem

if the firm is unsure about the trades union's intransigence. If its leaders are known to be pragmatic, backward induction would ensure the downfall of CS. However, in the face of uncertainty, the pragmatic leaders may wish to build, at least during the early stages, a reputation for intransigence and of unwavering commitment to the CS. This being common knowledge, management does not dare challenge for a number of negotiations, say $k < k'$, as doing so would give an incentive, even to pragmatic leaders, to put on a display of intransigence.

The higher the leader's initial reputation for intransigence the longer the first phase, during which the finite and indefinite horizon models are indistinguishable. At negotiation k' , however, the firm will challenge with positive probability. Faced by this threat to the credibility of its CS, the trades union will deliberate between meeting the challenge with a longer strike than the one predicted by

equation (2.12)

and acquiescing. Its equilibrium randomisation decision rule will be such that, in the event of a prolonged strike, its leader's reputation for intransigence will follow an optimal growth path. In the event this randomisation yields 'acquiescence', i.e. a settlement off the trades union's CS, the second phase is complete and the final phase begins, taking the firm and the trades union to the end of the horizon.

In summary, the second type of uncertainty was introduced into the analysis in order to salvage it from the claws of backward induction: uncertainty with regard to the trades union leader's intransigence. It is encouraging to note that the leader suddenly becomes central to the bargaining process, especially in view of the insight that even a pragmatic leader may wish to show signs of intransigence. Irrationality is, therefore, not necessary for the model to explain strike activity. Unlike models which rely on irrational behaviour (see Ashenfelter and Johnson, 1969; Cross, 1969), all we really need is that not everyone assumes

everyone else to be instrumentally rational. It is this *possibility* of intransigence that makes intransigence optimal even for rational, pragmatic bargainers who may even disdain bravado. Good trades unionists creatively exploit uncertainty surrounding their character and how to pick their moment for switching from a defiant to a conciliatory posture.

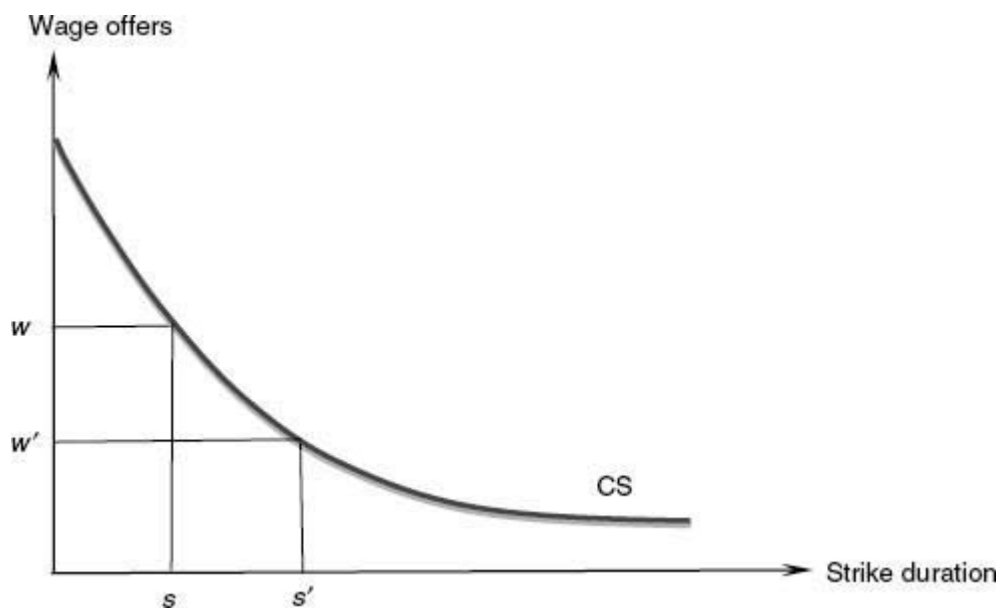


Figure 2.4

Employer resistance to wage demands.

2.2.7 The geometry of the third phase

The commencement of the third phase coincides with the first settlement of the trades union's announced CS. Thereafter unionists are expected to readjust the CS and the

firm embarks upon a search for the new CS.

[Figure 2.5\(a\)](#)

represents the set of possible candidates for a replacement of the original (now discredited) CS. Depending on a given prior of belief, management selects one of the available candidates as the new constraint in the maximisation of (

[2.9](#)

). If a strike does occur, there are two possible ways in which offers can be adjusted.

First,

[equation \(2.13\)](#)

may be activated if solidarity diverges from its anticipated path. Secondly, the employer may begin to swap one CS for another if trades union negotiators appear to follow a different CS path to the expected.

PROPOSITION 2.4 *With a CS reflecting a non-monotonic relation between wage demands and strike duration, employer offer paths will be characterised by hysteresis.*

Consider the first type of adjustment, where the firm is continually being surprised by unexpected rates of change in the size of the strike coalition. In the

case where σ is growing, the firm's estimate of its average profitability m (see

[Section 2.2.4](#)

for a definition) is falls and its isovalue map on the (w, s) plane becomes stepper. When expected profitability falls below m_0 – see

[Figure 2.5\(b\)](#)

– the sequence of employer offers undergoes a catastrophic jump. Conversely, if worker mobilisation (or solidarity) is rapidly dropping and m_0 is exceeded, the firm will reduce its offer discretely. The nature of the discontinuity will clearly depend upon the employer's ability to absorb information. If there is some delay between the reception of new information and the issuing of fresh offers,

[Figure 2.6](#)

takes over from

[Figure 2.3\(b\)](#)

and the discrete drop (rise) in the firm's offers [resulting from a continually increasing (decreasing) profit rate during the stoppage] will occur at an anticipated average in excess of (lower) than m_0 . When solidarity is growing, offers will follow the DEFA path. But when it is declining, a different path applies: ABCD; a case of *hysteresis*.

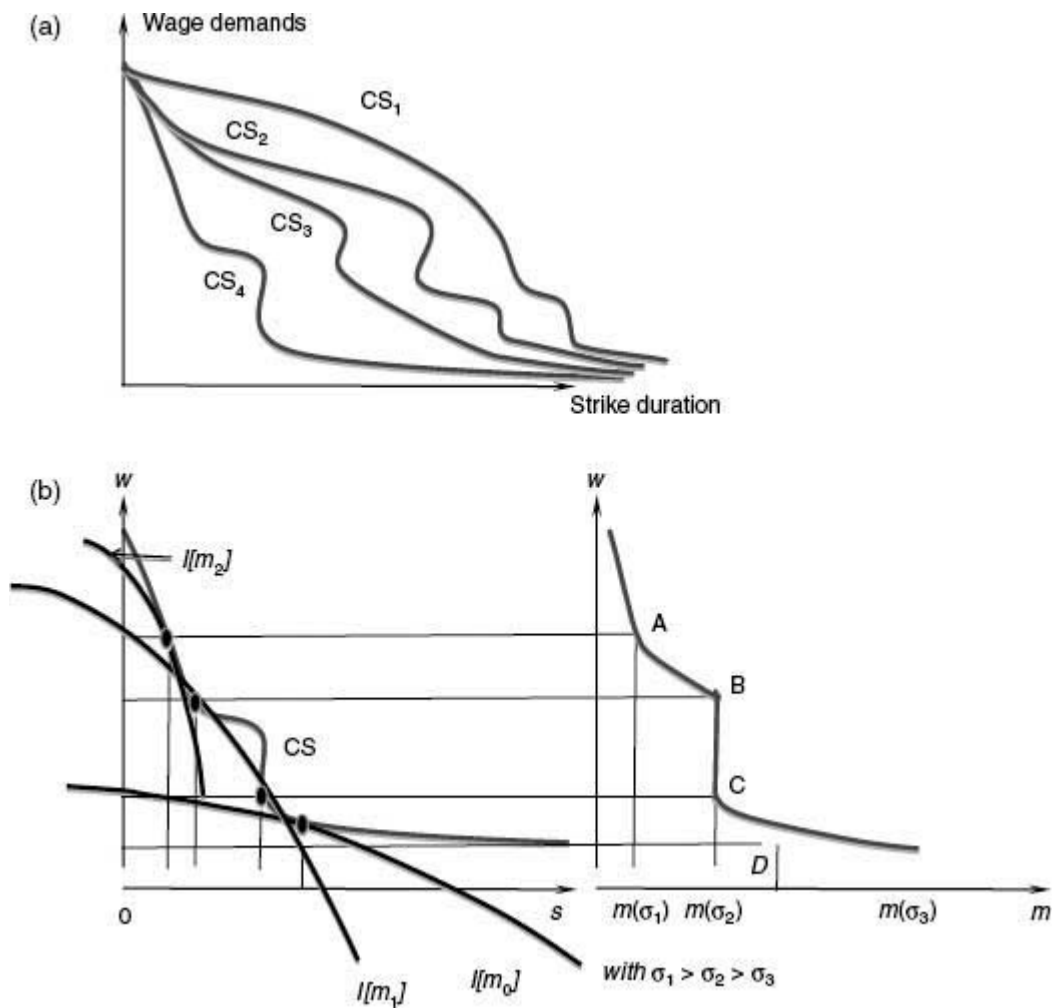


Figure 2.5

Non-differentiable employer resistance schedule.

Interestingly, for an average profit rate generated by strike-breakers in the region of m_0 , the finally agreed wage may differ profoundly, i.e. be equal to w_1 or to a

much lower w_2 , depending, for given average *per diem* mobilisation, on whether mobilisation has been rising or falling! As the same number of lot working days may lead to drastically different wage settlements, there is little doubt that the study of wage determination is incomplete without a thorough understanding of solidarity and its determinants.

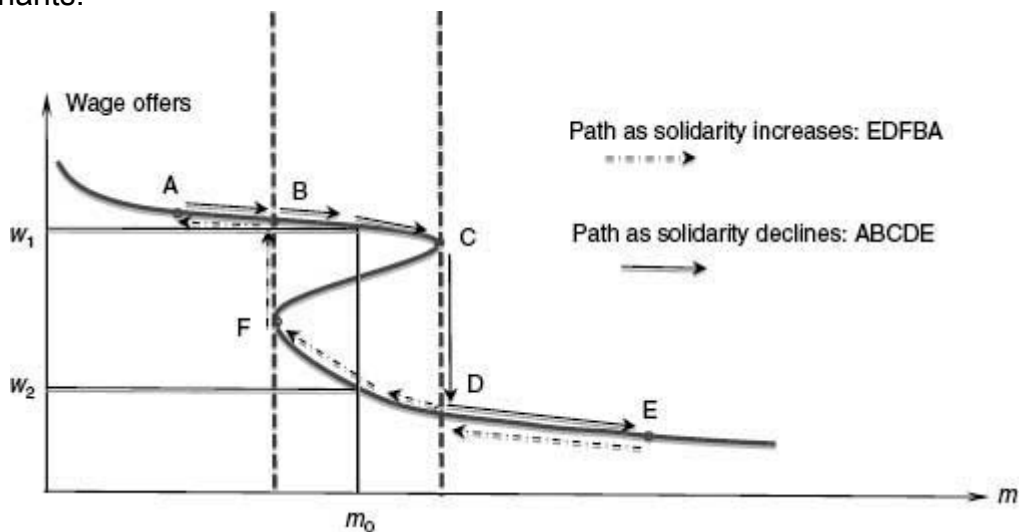


Figure 2.6

Hysteresis in the wage-mobilisation relation.

Let us now consider the second type of adjustment: if the firm is assumed to be well informed on m , fluctuations in its offers may still occur as managers are searching for the most likely CS. It is helpful to visualise the inter-play between the firm's fluid views on (i) the level of mobilisation and profits during the strike and (ii) the CS in terms of a three dimensional diagram featuring two control and one state variable. The state variable is the firm's wage offer which, at any point in time, is 'controlled' by the predicted lever of average profits courtesy of strike-breakers (i.e. m) and the precise location of the 'true' CS which is denoted by a parameter $\eta \in (0,1)$. As the employer moves from CS_1 to CS_η , η rises from zero to unity. Thom's classification theorem of elementary catastrophes applies

9

in the cases where an adjustment delay of the

Figure 2.6

type, suggesting that the

Figure 2.7

relationship in the most complicated one that can possibly emerge.

10

It is easy to see that

Figure 2.6

is contained within

Figure 2.7(a)

, as freezing η at some value, say η' , and varying m , yields the familiar paths ABCDE or EDFBA, depending on whether trades union solidarity-mobilisation is declining or increasing. The additional insight provided by the complexity of

Figure 2.7

becomes apparent when we consider the case of a relatively stable average profit level in the region of m_0 . If the firm takes as its constraint the CS with $\eta = \eta''$, its optimal wage offer to the trades union (w_M) corresponds to point M on the stationarity surface S . Supposing that the trades union appears to be more conciliatory than anticipated by η'' , a gradual upward revision of η from η'' to η' may lead to two qualitatively different paths, $\bar{\theta}_1$ and $\bar{\theta}_2$, for reasons that do not

extend beyond tiny, random perturbations. Note that both paths begin and end from/at precisely the same levels of mobilisation, expected strike costs (to the firm) and trades union concessions. Remarkably, they lead to two utterly different wage outcomes: w_N and w_D . A slightly smaller rate of mobilisation declines (during the dispute) may give rise to a substantially higher wage offer on behalf of the employer. The suspicion that the resolution of industrial conflict owes more to fortune than hitherto acknowledged may, after all, have an analytical foundation.

common cause of a very small number of workers.

2.2.8 Summary

The main point in the preceding pages is that anticipated worker solidarity and mobilisation plays a crucial role in determining the distribution of income between capital and labour. Indeed, I have shown that wages are not only determined by the degree of worker cohesion managers expect to confront if a strike occurs but, interestingly, they are also determined by solidarity's non-linear dynamics which matter over and above average mobilisation rates (e.g. its direction of change in mobilisation as well as perturbations around its path). If the process that allows trades unions to overcome the free riding tendencies of its members is of a dynamic nature and involves ideological and psychological externalities, the attitudes of workers during a dispute cannot be adequately summarised by a single, time-invariant trades union maximand or utility function.

In the model presented so far, we have seen how workers may come to the decision to transcend their prisoner's dilemma-like tendency to defect from the common cause as long as they draw socially contingent (net) utility from not crossing picket lines (in proportion to the degree of mobilisation that they observe). From this simple idea, a model of stable striking coalition formation was put together. Next, the trades union's strike call was rationalised as an exercise in forcing employers to come clean on the true capacity to pay higher wages. At that point, we studied the manner in which bounded horizon in the relationship between the union's leaders and the firm's managers creates a rich three-phase 'history' of negotiations. It was argued that the first two phases are either identical or similar to what would obtain if this firm (and its trades union) were to carry on forever, whereas the third phase (as we approach the end of, say, the union

leaders' term in office or a new CEO appointment) requires a major analytical departure. A tentative geometric investigation revealed important features of the effect of mobilisation and solidarity that the literature has consistently ignored.

In conclusion, the ability of a trades union to provide its members with an essentially public good is contained by the extent to which individual workers see the subjective (or psychological) components of their utility function dominate their monetary considerations. In commending the dynamic properties of the coalition formation process as a crucial missing link in the strike literature, the analysis formalises the labour movement's preoccupation with unity as a form of strength. It also points to the importance of ideology as a guarantor of the conventions which facilitate the convergence of common and private interest.

2.3 When negotiations involve staffing levels, it is hard to know what to aim at – for both managers and trades unions

11

2.3.1 Introduction

For at least seventy years economists have disagreed on the aims of trades unions, especially those concerning wages and staffing levels. Some have argued powerfully that unions trade jobs for wages, confining themselves to targets on the firm's labour demand curve (LDC).

12

Others, based on Wasily Leontief's famous 1946 article, have offered equally powerful arguments as to why unions will stray off the LDC, aiming at both higher wage and employment targets belonging to a contract curve.

13

Attempts at theoretical synthesis have added to the richness of the debate, with claims

that the LDC is the LCC

14

or that, under special conditions, the LCC tends towards the LDC.

15

Empirical evidence has failed to settle the debate, giving contradictory signals and suggesting that bargaining outcomes probably lie between the two loci.

16

Is there anything left to add to this literature? In this section I show that, despite the literature's impressive breadth, we seem to have missed out the problem's most important aspect: the effect of the level of rank and file mobilisation and solidarity on what managers and trades union negotiators choose as their bargaining aims. Indeed, by assuming that the union's sole constraint in satisfying its preferences on the wage/employment plane is the firm's resistance to labour's demands, we left out of the analysis perhaps the most crucial ingredient: the threat to the trades union posed by a waning propensity amongst its members to make private sacrifices for the trades union's collective interests. In what follows, I extend the model of the previous section (which dropped the implicit assumption of the workers' *unbounded* willingness to translate collective preferences to an intention to act collectively) to cases where the union cares not only about wages but also about jobs.

As we shall see below, when it is recognised that workers are susceptible to free-rider tendencies and do not *necessarily* mobilise in support of, and in proportion to, their aggregate preferences, a general theory of union targets emerges which accounts plausibly for the theoretical and empirical indeterminacy (i.e. targets falling neither on the LDC nor the LCC).

17

2.3.2 An overview of the algorithmic model

The aim of the model is to describe the algorithmic process by which trades union negotiators decide their wage and employment goals prior to negotiations with management. In this sense, it offers a theory of targets, as opposed to a theory of outcomes. According to it, rational negotiators ask: If we choose a certain target, (a) how much resistance should we expect from the employer, and (b) how much mobilisation in support of this target can we expect from members? Is such a mobilisation level sufficient to render this target feasible in view of (a)? And finally, of all the feasible targets, which one do we, the trades union leaders, favour?

At the outset, and in line with the existing literature, labour leaders are assumed to maximise a conventional aggregate utility function (W) on the wage (w) and employment (N) plane: $W = W(w, N)$. As an organisation, its end is improvements in pay and job security for its members. The level of potential mobilisation does not enter $W(\cdot)$ directly (since strikes are a means to an end) but only indirectly to the extent that, as all constraints, it helps determine the level of union utility.

18

By contrast, individual workers care about more than just wages and aggregate employment levels. As argued in

Section 2.2

, a worker's utility function $U(\cdot)$ features additional arguments: job security, self-esteem, psychological returns from a good standing within their community, expected wage losses during a strike etc. It is those concerns which will help explain, as they did in

Section 2.2

, why workers might be prepared to transcend the free-rider logic and form an intention to strike.

In arriving at a particular (w, N) target the trades union's leaders are assumed to employ the following algorithm:

Step 1. Select some bundle (w_i, N_i) and assess its merits as an appropriate target.

p 0

Ste Compute how much resistance (w_i, N_i) would encounter from the employer;
p 1 namely, what (credible) threat could motivate the firm to accept demand (w_i, N_i).

[Section 2.2.3](#)

answers the question thus: For target (w_i, N_i) to be imposed on the firm, the trades union must be able to muster combinations of threatened strike duration ($s > 0$) and worker mobilisation $\sigma \in (0,1)$ belonging to a region on the (s, σ) plane defined by some function $\Pi_i = \Pi(w_i, N_i, s, \sigma) > 0$ – see

[Section 2.2.3](#)

for the derivation of that function.

Ste Compute how much mobilisation, and for what length of time, labour leaders
p 2 can expect from their members in pursuit of target (w_i, N_i). In

[Section 2.2.4](#)

target (w_i, N_i) is shown to yield a set of feasible (s, σ) (i.e. duration-mobilisation) combinations. This set is delineated by function $s = \Omega_i(\sigma) [= \Omega(w_i, N_i, \sigma)]$. Subsequently the trades union's target (w_i, N_i) is defined as *feasible*, in

[Section 2.2.5](#)

, only if there exists at least one (s, σ) combination belonging simultaneously to Π_i (see Step 1) and Ω_i .

Ste Return to Step 1 and consider another target, say (w_k, N_k), in terms of Steps 1
p 3 and 2. Continue until all feasible targets have been identified. Then proceed to Step 4.

Ste Select the one feasible target which maximises the union's utility $W(w, N)$.

p 4

[Section 2.3.3](#)

analyses Step 1.

[Section 2.3.4](#)

offers the model of worker mobilisation necessary to activate Step 2 while

[Section 2.3.5](#)

completes the model by accounting analytically and diagrammatically for Steps 3 and 4. Finally,

[Section 2.3.6](#)

discusses the model's insights as well as weaknesses.

2.3.3 The constraint posed by the employer's resistance

The credibility of union leaders' plan to achieve target (w, N) by threatening a strike of s duration depends on whether this target will generate an expected mobilisation rate σ such that the firm would be at least as well off granting the union's (w, N) demand as it would be taking the strike.

19

In other words, union leaders believe that if the firm (a) expects the union's target (w, N) to have the potential of generating a strike with characteristics (s, σ^e), and (b) computes that the target/threat combination (w, N, σ, s) gives rise to non-negative expected net returns from granting labour's demand (i.e. $\Pi > 0$), then target (w, N) is feasible.

20

• Employer's expected returns from granting demands = $ER(\text{grant}) = \int_0^T [pQ(N) - wN]e^{-\delta t}$

• Employer's expected returns from taking a strike
 $ER(\text{take strike}) = \int_0^T C e^{-\delta t} + \int_0^s [pQ\{(1 - \sigma^e)N_0\} - w_0(1 - \sigma^e)N_0]e^{-\delta t} + \int_s^T [pQ(N_0) - w_0N]e^{-\delta t}$

where, p is the constant price of output, $Q(\cdot)$ is the firm's output as a function of employment, N_0 is the number of workers employed initially, w_0 is the previous wage (which strike-breakers are paid during a dispute), C is a lump sum loss due to the strike (additional to lost revenues; e.g. bad publicity, lost good will), δ is the firm's discount rate, T is the contract's duration and, finally, w and N are the union's wage and employment targets (or demands).

Thus, inequality (

2.14

) is the employers' decision rule on whether to grant the trades union's demand or choose to take a strike in a bid to deflate it.

Management's Decision Rule: Grant initial demands iff

$$\Pi = \text{ER}(\text{grant}) - \text{ER}(\text{take strike}) \geq 0 \quad (2.14)$$

Assuming that the firm's resistance to a trades union target is proportional to the value of Π (i.e. the net expected gains from acquiescence), the firm's propensity to take a strike will be identical for targets generating the same Π .

These contours can be shown to be downward sloping and concave to the origin under fairly general conditions.

21

Their slope is actually given by:

$$\frac{(1 - e^{(-\delta s)})}{(\delta e^{(-\delta s)})}$$

$$\times N_0 \frac{\text{Strike breakers' wage minus their marginal revenue product}}{(\text{Value of lost output}) + (\text{Fixed strike costs}) - (\text{Wages saved during strike})}$$

2.3.4 The constraint posed by imperfect worker mobilisation

To strike or not to strike? As in

Section 2.2.2

, so too here workers are assumed to derive utility additively from two sources: expected earnings during the life of the contract under negotiation (T), including any period of industrial conflict, and from non-pecuniary, psychological factors relating to their stance during a strike. They expect a decision to strike to affect their income by altering their prospects for keeping their job (e.g. strikers run a higher risk of dismissal) and by incurring wage losses during the dispute.

22

On the other hand, they anticipate a psychological cost in case they break the strike which may take either an intrinsic form (e.g. feelings of guilt for having betrayed one's colleagues; perhaps of being morally defective) or a purely reputational form (e.g. not wanting to be seen to cross picket lines for fear of becoming an outcast). Either way, breaking a strike creates a certain image (perhaps a self-image) that affects one's utility for a long time.

More precisely, the individual worker's expected utility is given by

2.15

and

2.16

below depending on whether one has formed the intention of heeding one's union's strike call or not.

$$U^J = a_1 \left\{ \int_s^T \Lambda w e^{-rt} dt + \int_0^s b e^{-rt} dt \right\} + a_2 \left\{ \int_s^T R_J e^{-rt} dt \right\} \quad (2.15)$$

$$U^D = a_1 \left\{ \int_s^T M w e^{-rt} dt + \int_0^s w_0 e^{-rt} dt \right\} + a_2 \left\{ \int_s^T R_D e^{-rt} dt \right\} \quad (2.16)$$

where superscripts J and D denote decisions to 'join in' the strike or to 'defect'

respectively, the α 's capture the marginal utilities from anticipated pecuniary and non-pecuniary inter-temporal benefits, and $t = 0$ is the moment in the future when the union is to take its announced wage/employment target to the employers. The negotiations are assumed to be instantaneous, in which case rejection of the target/demand leads to a strike. Strike duration s is the approximate (or average) length of time the union's leaders have asked the workers to prepare for in support of the collective target, \square and M are the probabilities of remaining in employment for strikers and non-strikers respectively, w is the union's wage target,

23

r is the worker's discount rate, b is the per period strike benefit payable by the trades union to strikers, w_0 is the previous wage rate (which is also the wage strikebreakers receive) and, finally, R is the magnitude of the psychological effect (or reputation) that the worker's decision will engender (R_j if one decides to join, R_d otherwise).

Assuming that workers and their leaders have common knowledge of the above variables and parameters, our worker will intend to strike depending on the rule below – which you will notice is the same as that of

equation (2.1)

of the earlier model in

Section 2.2.2

:

	join	if $\rho\gamma > z$
<i>Decision Rule</i>	strike break	if $\rho\gamma = z$
	indifferent	if $\rho\gamma = z$

where $\rho = R_j - R_d$ is the *psychological/reputational/moral loss* from strike-breaking
e

$\gamma = \alpha_2/\alpha_1$ is the relative utility valuation of such a loss

$z = \lambda w[\exp(-rs) - \exp(-rT)] + (w_0 - b)[1 - \exp(-rs)]$ is the sum of the inter-temporal utility valuation of the *extra job insecurity* and the *lost wages* facing those workers who choose to strike

$\gamma = M - \square$ is the increase in the probability of being laid off for workers who join the strike.

In plain language, a worker will form the intention of joining a strike in pursuit of a particular wage-employment combination (w , N) provided the moral, or reputational, loss from refusing to cross picket lines exceeds (i) the gains in job security from breaking the strike and (b) the lost wages during the strike. Note that, as before, the only factor in this new version of rule (

2.1

) mitigating against a paralysing free rider problem is function ρ . This shows that the prospects of collective action are very loosely connected to the collective purpose of a potential strike. Indeed, because the negotiated outcome is a form of public good (to be enjoyed by strikers and strike-breakers alike), the union's wage and employment targets do not affect the individual worker's pecuniary losses from striking [i.e. function $z(\cdot)$], except insofar as employers have the opportunity to penalise strikers.

24

As argued in

Section 2.2.2

, and depicted in

Figure 2.1

, a natural extension of the thought that workers care about the moral or ideological content of their actions is that such rewards depend on the context and one important feature of the social context is what others are doing so that workers adapt to the observed group intentions. In short, the value of ρ above is an increasing function of σ or, in simple terms, the shame, guilt or worry about one's image from (or indeed the excited anticipation of) crossing picket lines is usually highly sensitive to the proportion of those who also cross picket lines. A strike where only a lone radical keeps a sad vigil by the factory gates must surely be less stressful emotionally to break than one in which most of one's colleagues implore one to stay out. Thus I posit again a positive relation between ρ and σ ; the *solidarity function* of

equation (2.4)

However, it is not only the non-pecuniary component on the l.h.s. of (

2.4

) above which depends on the mobilisation rate σ . For example, workers anticipate that, if everyone is to walk out during a dispute the chances of being fired *because* one will strike are much smaller than in a situation where only a tiny minority of

the workforce plan to walk out. Thus loss function $z(\cdot)$ also varies with σ via the latter's effect on λ or indeed on $(w_0 - b)$.

25

In summary, the following functions are defined:

ρ = The *solidarity function* which links non-pecuniary gains from striking with $\rho(w, N, \sigma)$; the *mobilisation rate* as well as with the union's targets

σ ;

$\partial\rho/\partial\sigma$

> 0

z = The *loss function* which captures the potential losses being fired if one strikes plus the loss of wages during a dispute

$z[\lambda(\sigma)w, s, \sigma]$

γ_i \square where $\Gamma(\cdot)$ is the distribution of private valuations of non-pecuniary gains from participating in collective action relative to the pecuniary losses; γ_m and γ_d being the mean and standard deviation of that distribution.

As inequality (

2.4

) may hold for some but not all *mobilisation rates*, the next step is to describe the two types of potential mobilisation equilibria. The first type involves corner solutions engendering unity equilibria (since whenever mobilisation occurs it is perfect) while the second type makes interior solutions possible. The latter are referred to as disunity equilibria since they feature a stable blend of strikers and strike-breakers. These two cases are equivalent to the two sets of equilibria in

Figure 2.2

, depending on the relative slope of the ρ and z functions in rule 2.1. The only difference is that here the z function is significantly more complex, as it takes account of the probability of remaining in employment after the industrial dispute is over.

$$\text{Unity Equilibria: } \gamma_m[\partial\rho/\partial\sigma] \geq \frac{\partial z}{\partial\sigma} \quad (2.17)$$

Under inequality (

2.17

), and provided even the least 'ideological' worker would be unwilling to cross the picket

lines when everyone else stays out,

26

it is easy to show that, in equilibrium, either all workers will walk out ($\sigma = 1$) or none will ($\sigma = 0$). Union members will either fight disputes united or not at all. To see this, define σ^* as the mobilisation level that solves (

2.17

) as an equation. Now consider the simplistic case where workers are identical (i.e. $\gamma_d = 0$). If each worker expects mobilisation to exceed threshold σ^* , then clearly inequality (4) will apply for each and thus everyone will be expecting to strike. Otherwise the prospects of mobilisation are doomed.

27

Next, consider the more general case in which workers differ amongst each other in two important ways. Firstly, they have different relative valuations of reputation or of job security or of lost wages (i.e. $\gamma_d > 0$), possibly due to ideological differences or family circumstances; and, secondly, different degrees of pessimism (or optimism) about the likelihood of successful collective action. To model the latter, let q_i be the i th worker's subjective probabilistic expectation

that all workers will heed their union's strike call for s periods (i.e. the probabilistic expectation that $\sigma = 1$). Looking at (

2.17

) again, whenever $\sigma^* > 0$ (or $\sigma^* < 1$), perfect mobilisation is guaranteed (or it is doomed). Rational workers will therefore form q_i beliefs accordingly but may be also influenced by a private predisposition towards pessimism [reflected in (

2.18

) below by parameter η_i].

$$q_i = 1 - \eta_i \sigma^*; \quad \text{where } \eta_i \sim N^\tau(\eta_m, \eta_d^2) \quad (2.18)$$

28

Because unity equilibria involve bandwagon effects which lead to either perfect or zero worker mobilisation, each worker will form the intention of joining the announced strike with probability q_i . Thus the expected rate of mobilisation σ^e is the same as the mean of q_i which, following (

2.18

), equals $1 - \eta_m \sigma^*$. Confirming the importance of optimism, viz. the union's capacity to mobilise, the expected degree of worker mobilisation will equal the probability of perfect mobilisation and will be a decreasing function of average pessimism amongst the workforce:

29

$$\sigma^e \equiv \Pr(\sigma > \sigma^*) = 1 - \Phi^\tau[((1 + \eta_m)\sigma^* - 1)/\eta_d] \quad (2.19)$$

$$\text{Disunity Equilibria: } \gamma_m \left[\frac{\partial \rho}{\partial \sigma} \right] \geq \frac{\partial z}{\partial \sigma} \quad (2.20)$$

Under condition (

2.20

), interior mobilisation equilibria emerge, similar to those in

Figure 2.2

when the slope of the z function is positive. Then, a stable proportion of workers will form the intention of striking while the rest intend to continue working. We call these **disunity equilibria**. When disunity equilibria [i.e. $0 < \sigma^* < 1$] beckon, unions must prepare for imperfectly supported strikes with the marginal worker indifferent between crossing picket lines and joining them [i.e. worker k for whom $\gamma_k \rho(\sigma^*) = z(\sigma^*)$].

30

So far we have seen that expected mobilisation (σ^e) depends on both socially determined and private factors. Importantly, different union (w, N) targets may lead to

different levels of anticipated mobilisation quite *independently of the worker's private preferences on the w - N plane*. This is so because mobilisation depends largely on the social conventions within the union, as reflected in the *solidarity function* $\rho(\sigma)$. Thus different ρ functions will yield different mobilisation equilibria for the same private wage and employment preferences. Competent union leaders will keep a keen eye on the dynamic path of solidarity in order to avoid threatening employers with harmless strikes involving a minuscule number of the 'usual trouble-makers'.

31

To do so they must assess every wage and employment target with the prospective mobilisation rate in mind – see Step 2 in the algorithm sketched out in

Section 2.3.1

Once they have worked out how much employer resistance to expect for each potential (w, N) target, they must establish the corresponding bundles of (s, σ^e) which their members would, potentially, bring on in support of that target. Naturally there are more than one ways a union can achieve a certain (feasible) target: A long but poorly supported campaign might be equally effective as a short, sharp strike (as long as both fall on the same Π contour in

Figure 2.8(c)

; see

equation (2.14)

for the derivation of these contours).

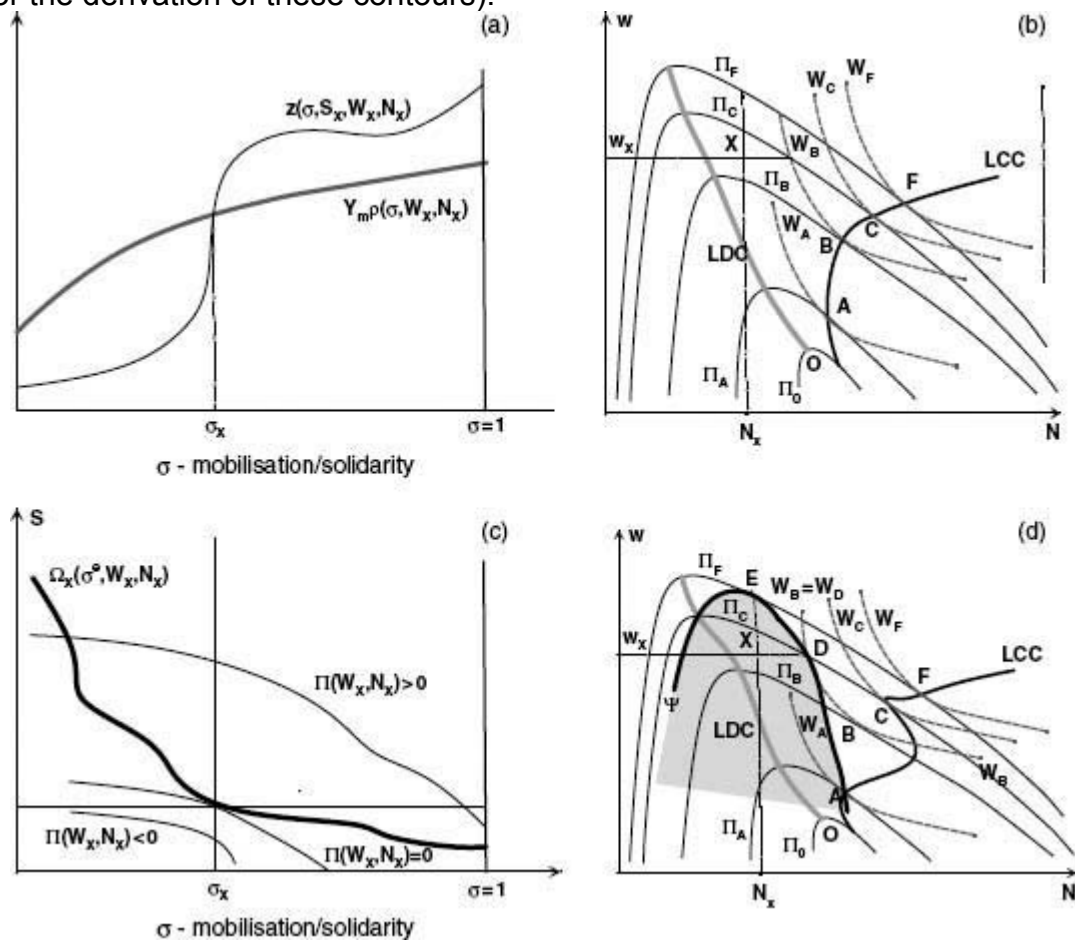


Figure 2.8

The determination of the trades union's constraint (or boundary targets).

In general, for each target the union will face a trade-off between strike duration and expected mobilisation (the *duration vs. mobilisation frontier* hereafter) given by (

2.21

) below and derived as follows: In the case of **unity equilibria** [i.e. under inequality (

2.17

)] the inverse relationship ω between s and σ^e is mediated by equation (2.19)

whereas in the case of **disunity equilibria** [i.e. under inequality (

2.18

)] the trade-off is given by

equation (2.21)

32

$$\text{Letting } \sigma^e = \omega(s) = \begin{cases} \sigma^*(s, w, N) \text{ under inequality (2.20)} \\ 1 - \Phi_\tau \left[\frac{(1 + \eta_m)\sigma^*(s, w, N) - 1}{\eta_d} \right] \\ \text{under inequality (2.17)} \end{cases}$$

then, in general, the union's *duration vs. mobilisation frontier* is given by

$$s = \Omega(\sigma^e) = \omega^{-1}(\cdot) \text{ with } d\Omega/d\sigma < 0 \quad (2.21)$$

Equation (2.21)

fulfils an important part of Step 2 (see

Section 2.3.1

), namely the computation of the degree of mobilisation that each target will engender. In the next section Step 2 will be completed and Steps 3 and 4 will be addressed.

2.3.5 The trades union's optimal target

Union leaders are now ready to establish whether some target $(w, N)_i$ is feasible.

Section 2.3.3

tells them that the employer will have cause to consider $(w, N)_i$ seriously if the firm fears a strike characterised by some duration and mobilisation (s, σ^e) combination such that $\Pi_i = \Pi[(w, N)_i]$, i.e. the l.h.s. of (

2.14

), is non-negative. Can they rely on workers to generate such a threat? The previous section, and in particular

equation (2.21)

, answers their question. From

equation (2.21)

they know that target $(w, N)_i$ will engender potential combinations (s, σ^e) satisfying $s = \Omega_i(\sigma^e) = \Omega[\sigma^e, (w, N)_i]$. So, is (w, N) feasible? Yes, provided there exists at least one combination (s, σ^e) belonging simultaneously to $\Pi_i \geq 0$ and to $s = \Omega_i(\sigma^e)$. Otherwise (w, N) cannot be a believable target. [Diagrammatically this condition requires that in

Figure 2.8(c)

there is at least one common point between a Π contour (which involves a non-negative Π value) and the Ω frontier.] The formal definition of feasible targets follows:

Feasible targets: If the trades union can select a target $(w, N)_u$ which engenders a potential mobilisation frontier $s = [\Omega\sigma^e, (w, N)_u]$, of which at least one combination $(s, \sigma^e)_u$ is such that the firm's inter-temporal profit from refusing to grant $(w, N)_u$ [choosing instead to take a strike of $(s, \sigma^e)_u$ duration and degree of mobilisation] is at least equal to its profit from granting $(w, N)_u$ without industrial conflict, i.e. if $\Pi[(s, \sigma^e)_u, (w_0, N_0), (w, N)] \geq 0$, then $(w, N)_u$ is a *feasible target*.

Note how the feasibility of each potential union target is partly determined by its reception by the rank and file. Therefore, unlike the existing literature in which the union

leaders' relative bargaining power is exogenous to their choice of target, in this model their power to withhold labour, and thus to bargain, is endogenous. *Ex post* for each feasible target $(w, N)_u$, and if the firm's per period profit are given by function $\pi(\cdot)$, labour's bargaining power can be computed by (

2.22

):

Index of union bargaining power:

$$\Theta\{(w, N)_u\} = [\pi(w_0, N_0) - \pi\{(w, N)_u\}] / \pi(w_0, N_0) \quad (2.22)$$

where Θ denotes the rents a union can wrestle from firms by adopting target $(w, N)_u$ as a proportion of the firm's rents during the life of the previous contract.

Rational union leaders adopt targets such that the subsequent level of Θ enables them to attain the maximum aggregate union utility $[W(w, N)]$. To this end, they will want their target not only to be *feasible* but also *effective*, in the sense that it is underpinned by the most potent combination of potential strike duration and mobilisation. Amongst their *effective* targets they will then select those (defined below as *boundary targets*) which yield the best deal on the wage front given the negotiated employment level (and vice versa). Finally their optimal target will be

the boundary target which yields maximum union utility $W(\cdot)$. This process is delineated in detail below:

Effective targets: A *feasible target* $(w, N)_u N'$ is also an *effective target* if, for the same *feasible* wage and employment demands, the union cannot boost its *bargaining power* Θ simply by substituting some strike duration with greater worker mobilisation (or vice versa). Thus for $(w, N)_u N'$ to be effective, then on the (s, σ) plane one of the firm's iso-resistance (or isoprofit) Π curves [derived in

Section 2.3.3

and corresponding to target $(w, N)_u N'$ must be tangential to the *duration vs. mobilisation frontier* [i.e. $s = \{\Omega \sigma^e, (w, N)_u N'\}$] (which was derived in

Section 2.3.3

) – see

Figure 2.8(c)

for an example.

Boundary targets: An *effective target* $(w, N)_b$ is also a *boundary target* if w_b maximises Θ given the employment target N_b and, simultaneously, N_b maximises Θ given the wage target w_b . Thus beginning with a *boundary target*, even a minuscule increase in the wage target above w_b (while holding the employment component of the target, N_b , constant), or in the employment target (while holding the wage component of the target, w_b , constant), would render the target infeasible (and thus ineffective). Let Ψ be the locus of all *boundary targets*, and ψ be the concave (to the origin) *boundary* of Ψ . Then, assuming convex union preferences $W(\cdot)$, ψ is the *boundary* containing all wage and employment targets that a rational union would consider.

Maximal and optimal targets: The union's *maximal target* is the (w, N) bundle belonging to ψ which forces the firm to yield its maximum concession; that is, bring its profit rate to the minimum level below which the firm would rather cease production permanently. Of course, only accidentally will the union prefer to adopt its *maximal target* over other targets within its *boundary* ψ .

33

In general, the union will select the target in ψ which maximises the union's utility $W(w, N)$. The target which does this is, naturally, the union's *optimal target*.

2.3.6 What does this all mean regarding the trades union's aims?

A geometrical analysis of the trades union's optimal targets

Figure 2.8

captures the essence of the above target selection model.

Figure 2.8(b)

replicates the standard treatment of wage and employment targets; namely, by means of the LDC and Leontief's contract curve (LCC).

Figure 2.8(a)

and

Figure 2.8(c)

reflect the ways in which the union expects its choice of target to affect, respectively, worker mobilisation (see

Section 2.3

) and the firm's level of resistance (see

Section 2.3.3

). To illustrate an *effective* union target, consider target X in

Figure 2.8(b)

. Is it *effective*? Yes, provided the curves in parts

Figure 2.8(a)

and

Figure 2.8(c)

apply when X is the union's chosen target.

Relation $\Pi(w_x, N_x) = 0$ in

Figure 2.8(c)

– see

Section 2.3.3

, inequality (

2.14

) for its derivation – depicts the combinations of strike duration (s) and expected mobilisation levels (σ^e) which would cause the firm to be indifferent between a strike and acquiescence to the union demand (w_x, N_x). On the other hand, relation $\Omega(\cdot)$ – see

equation (2.11)

for its derivation – captures the combinations

of s and σ^e with which workers empower their leaders *given that the union has adopted target X*.

Figure 2.8(a)

completes this scenario according to which X is an *effective* target. We can confirm this by inspection of: (i)

Figure 2.8(c)

reports that target X gives rise to a single combination

34

(i.e. the tangency point between Π and Ω) capable of overcoming employer resistance ($s = s_x, \sigma^e = \sigma^x$); and (ii)

Figure 2.8(a)

which shows that when workers hear of target (w_x, N_x) and of the fact that they may be called upon to strike for s_x periods, then the equilibrium

35

mobilisation rate is precisely σ^x . In this sense, X is a target which the union can aim for realistically, even if only just.

However, the fact that X is *effective* does not necessarily mean that it is also a *boundary target*. Suppose, for example, that the union were to retain employment target N_x but increase its wage target beyond w_x [i.e. move in the direction of E – see

Figure 2.8(d)

]. What effect would this have on its two constraints (i.e. the one posed by members and the one presented by the employers)? In

Figure 2.8(c)

the firm's iso-resistance (Π) curves would shift upwards, since the union's higher claim would cause greater resistance. To render the new target feasible, the Ω function would also have to shift upwards; that is, workers would have to be willing to strike for longer, or to strike-break less, or both.

But would they do so? Perhaps, though not necessarily. For example, if the union now asks its members to be ready to strike for longer, s will rise above s_x and the $z(\cdot)$ loss function would shift upwards, thus *ceteris paribus* reducing the degree of expected mobilisation σ . However, if the solidarity function $\rho(\cdot)$ were to rise even faster than $z(\cdot)$ as a result of the higher wage target, then perhaps the necessary level of σ will be produced, in which case the new target will be feasible [I shall return below to possible causes of $(\partial\rho/\partial w) > (\partial z/\partial w)$]. In other words, as the union's target shifts from X towards E, it will remain feasible as long as the union's capacity to inflict strike costs on the firm rises faster than the firm's resistance. Similarly, if the union were to increase its employment target *ceteris paribus* (i.e. in the direction of D) again its power would have to increase sufficiently to overcome the firm's resistance.

36

Of course there is a limit (which we defined as ψ , the union's *boundary*) beyond which greater ambition leads to infeasibility as the rate of increase in employer's resistance will overwhelm the rate of increase in mobilisation; i.e. once the *boundary* is reached and (

2.14

) ceases to hold. At that *boundary* target [e.g. points E and D in

Figure 2.8(d)

] the union's bargaining power will have been maximised given either its wage or its employment target. In the context of

Figure 2.8(d)

, clearly the rational union choice of target is D as it represents a tangency point between its *boundary* ψ and one of its indifference curves (W_b). [Note that E is the maximal target; i.e. the one which would, if adopted, minimise the firm's profit].

Three reasons why the optimal target will not lie on the LCC

If perfect worker mobilisation could be guaranteed, then it would be independent of the union's choice of target, in which case the union's optimal target would always fall on LCC. However, if perfect mobilisation cannot be assumed and, instead, depends on the target which workers are called to mobilise in favour of, the *optimal* target [see

Figure 2.8(d)

] will lie off the LCC (excepting very restrictive circumstances). Why? To rephrase the question, why is the union's capacity to extract rents from the firm (i.e. its bargaining power Θ) not maximised by a target on the LCC?

37

The simple answer is that the LCC reflects, as in the rest of the literature, the preferences over the wage and employment space of (a) the employers and (b) the trades' union. However the workers' decision to join a strike is only loosely connected with these preferences, even if the union's indifference curves (W) in

Figure 2.8(d)

splendidly aggregate individual workers' valuations of expected income from different union targets. Indeed this is no more than a restatement of the well known result (see Olson, 1965) that individual contribution to collective action cannot be explained in terms of the individual's pecuniary benefits from the latter. Once the problem of translating collective preferences into collective action is addressed, it ought to surprise no one that the union's *optimal* target will only lie on the LCC under special circumstances. Three reasons for this are offered below:

REASON 1: THE UNION'S HISTORY AND SOCIAL FABRIC

The formation of any union, indeed union membership itself, implies that workers have overcome an urge to free ride on their colleagues' collective action. How does this happen? The answer must surely involve customs, reputation, solidarity, fear of ostracism etc. In this chapter all these are encapsulated in the union's *solidarity function* $\rho(\cdot)$ which appears to the individual worker as an 'externality'. This function is specific to each particular union or group of workers and has conceivably evolved as part of the same evolutionary process that led to the establishment and survival of the union.

Consider the following example. A trades union which came into being through campaigns to raise the wage to some level considered by workers as 'fair' [perhaps in the spirit of Akerlof, 1982], may yield a function ρ which is more sensitive to wage targets consistent with the old aims of wage justice. Thus, while a worker may privately prefer a wage of \$300 per week (and, say, more job security), she may be less willing to cross picket lines set up to demand \$350 per week than if the demand was for \$300 per week! Why? Because if there is a social convention within the union or community according to which \$350 is a fair wage worth fighting for [that is, $\rho(w = \$350) > \rho(w = \$300)$ for all values of σ],

38

breaking the strike in support of the \$350 claim is more painful.

39

Thus, one reason why the union's target cannot be assumed to belong to the LCC is that the latter does not take into account such social norms, which nevertheless play a crucial role in fashioning the rate of mobilisation and, thus, union bargaining power.

40

Returning to the union targets in

Figure 2.8(d)

, we can now see that, unless those targets which are most 'painful' for workers to break (relative to the 'pain'

from greater risk of losing one's job or from wage losses during the strike) are also the ones that maximise aggregate utility for each level of employer resistance (or profit), the union's *maximal* and/or *optimal* targets will lie off the LCC.

41

Geometrically speaking, the above example suggests that, while pondering their best LCC target [i.e. target A in

Figure 2.8(d)

], trades union leaders will realise that, if they push their target up the ψ boundary they will be increasing both union utility and bargaining power (Θ) as the mobilisation rate rises for each level of strike duration. Of course, union leaders prefer outcome C to D. But even though targets C and D would generate the same degree of employer resistance if adopted by the union, C (unlike D) would fail to spawn the level of mobilisation that would render it feasible.

REASON 2: INSIDERS ARE POTENTIAL STRIKERS; OUTSIDERS ARE NOT

As the literature has traditionally recognised (except perhaps Oswald, 1994), a union's maximand embodies the preferences not only of working members but also of (at least some) unemployed ones. This presents us with a second intuitive explanation why the LCC is not the locus of union targets: for it is only the actions (and readiness to mobilise) of the employed members which determines the union's bargaining power (and thus *boundary* ψ).

Consider the simplest variant in which the wage equals w_0 and union members face an identical probability of being fired or hired proportionate to fluctuations in employment which currently stands at N_0 . The union is tossing up two different targets which would be fought with equal rigour by the employer (i.e. they correspond to the same level of Π): a wage claim $w_1 > w_0$ (with employment constant at N_0) or extra jobs $N_1 > N_0$ (with the wage constant at w_0). Suppose further that a majority of union

members prefer the latter target and so $W(w_0, N_1) > W(w_1, N_0)$. Conventional bargaining models predict that the union would opt for target (w_0, N_1) . Not necessarily so in this model: For if the working members are significantly more likely to mobilise in support of target (w_1, N_0) than of (w_0, N_1) , the latter will not be adopted by the union leaders, not because they attach more importance to the views of insiders (or to the views of the median voter, as Oswald, 1994, would have it) but simply because their preferred target (w_0, N_1) is infeasible. Note that, in juxtaposition to various complex arguments in the outsider-insider literature, the simple thought here is that insiders affect the union target directly due to their perfect monopoly over the capacity to walk out during a strike.

In summary, even in cases where $W(w_0, N_1) > W(w_1, N_0)$ and $\pi(w_0, N_1) = \pi(w_1, N_0)$, the trades union may still select target (w_1, N_0) – rather than (w_0, N_1) – provided the *solidarity function* amongst working members is more responsive to wage claims than to job issues; that is, if

$$\left. \frac{\partial \rho}{\partial w} \right|_{w=w_0, N=N_0} > \left. \frac{\partial \rho}{\partial N} \right|_{w=w_0, N=N_0}$$

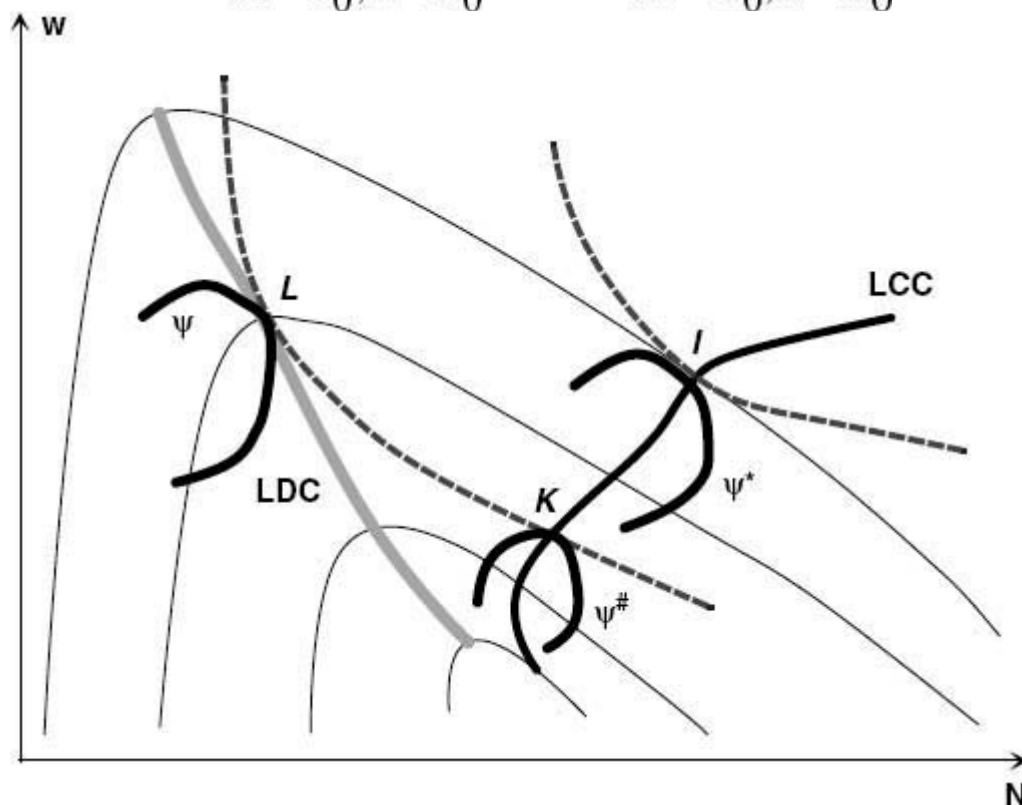


Figure 2.9

The effect on targets of asymmetric power to stop productions.

Of course, this is not to say that the insiders' monopoly over mobilisation will necessarily lead to targets above and on the left of LCC [see

Figure 2.8(d)

]. For example, the insiders may privately prefer a higher wage but still find it harder to cross picket lines when the union campaigns for jobs on the behalf of their unemployed colleagues. This may yield an *optimal target* below and on the right of the LCC; as in

Figure 2.9

where the union's *boundary* $\psi^\#$ leads to the target K .

In general, the only occasion when the *optimal target* will fall on the LCC is when working members find it hardest to break strikes in support of a union wage/employment target which happens to maximise aggregate utility for a given employer resistance (or

profit) level.

Figure 2.9

illustrates this special case, where the union's boundary ψ^* is simultaneously tangential to union indifference and an employer isoprofit curve at target I.

REASON 3: WORKER HETEROGENEITY

So far workers have been assumed to be identical in all respects except (a) their relative valuation of non-pecuniary gains from striking (the γ s) and (b) their optimism with regard to the *mobilisation rate* (the ζ s). Let us consider the effect of additional sources of heterogeneity. Suppose there is a group of workers whose chances of being fired if they join a strike increase by far less than the rest (i.e. they enjoy a lower λ than others for each union target); e.g. a group of workers whose skills are in relatively short supply. Their decision to mobilise will clearly be affected far less by the fear of being fired for joining a strike than other workers.

They thus constitute a group of workers the union must appeal to before planning any campaign. Moreover, it is often the case that such skilled workers:

- (i) enjoy a disproportionate capacity to inflict costs on the employer by, for example,
 - i) flicking some switch and instantly interrupting the production line
- (ii) are often linked to each other with group specific conventions: they are far more
 - i) interested in whether other members of their group mobilise in favour of the union campaign than in whether other workers do (i.e. the solidarity function ρ amongst these workers is quite separate from every one else's).

From the preceding analysis it should not be hard to demonstrate that when (i) and (ii) hold, the union's *boundary* – and therefore its *optimal target* – has acquired a hitherto hidden determinant: the *degree of control over the production process* enjoyed by sub-groups of workers which are bound together not only by union-wide conventions but also by sub-group specific social customs and norms. Indeed, it should be the case that, *ceteris paribus*, changes in the degree of control over production by different groups of workers can cause large shifts in the union's *optimal target*.

Worker heterogeneity gives us the opportunity to add another special case to

Figure 2.9

(in addition to LCC targets such as I). Suppose, for instance, that there is a group of workers who enjoy an almost unlimited capacity to interrupt production, a low probability of being fired if they strike, and no concern for crossing picket lines unless the strike is over wages. Then the union's *boundary* may be similar to ψ , in

Figure 2.9

, and the trades union will be constrained to a target (e.g. L) close to the LDC. The reader will notice the difference between this explanation of LDC targets and that in Oswald (1994): Whereas the latter postulates horizontal union indifference curves in

Figure 2.9

(implying that leaders adopt the median voter's preference for a higher wage without any concern for more jobs), this model makes no such assumption. LDC targets are selected even though the trades union cares for employment (witness the downward sloping indifference curve through target L in

Figure 2.9

). Why? Because they are constrained by that group of workers which will not strike otherwise and whose support is essential, given their location in the production process.

In summary, one appealing feature of this model is that it encompasses existing theories as special cases without needing to tell implausible stories about the shape of the trades union's indifference map. Indeed, the model can account for fluctuations in a union's targets (ranging from point K to target L in

Figure 2.9

) while all along the wage and employment preferences of each worker, the union leaders, and the firm's managers remain constant!

2.3.7 Three potential criticisms

Criticism 1: 'the optimal target is unstable'

Suppose target D in

[Figure 2.8\(d\)](#)

is achieved. Do union leaders not have an incentive to negotiate with management a new bargain somewhere along segment BC of the LCC? If such a Pareto improvement is taken up, do we not return to the LCC – i.e. to Leontief's 'efficient bargaining' model? Naturally. But then again rational leaders will know that, if they mislead the rank and file in this manner no announced (w, N) target off the LCC will be credible in future, forcing them thereafter to the lower utility $W(\cdot)$ level associated with targets such as A (instead of D). As long as leaders care about future outcomes, they will strive to maintain the credibility of their pronouncements and refrain from renegotiating *optimal targets*.

42

Criticism 2: 'factors affecting mobilisation should be included in the union's maximand'

Is it necessary, one may ask, to channel the new determinants of workers' actions through the union's constraints? Can we not simply write factors, such as the responsiveness of the *solidarity function* ρ to changes in the union targets, into the union's aggregate preferences and, by so doing, retain the analysis employed in the 'efficient bargaining' story (i.e. rather than abandon the LCC simply relocate it)? The answer is that we cannot do so while retaining the intuitively appealing dynamic aspects of the mobilisation model in

[Section 2.3.4](#)

. For example, the importance of bandwagon effects and social norms in this chapter would be lost entirely if they were subsumed in some aggregate function, for reasons known well since at least Akerlof (1980).

Furthermore, we should not want to over-burden the union's maximand for another reason: When worker heterogeneity was discussed above, it became clear that union targets depended on more than just workers' preferences. For instance, the organisation of production also appeared as a determinant. If we were to introduce such factors through the trades union's objective function, $W(\cdot)$, rather than its constraint – and mindful of the fact that the employer has the capacity to reorganise the production process – we would effectively grant one bargaining side (i.e. the firm) direct control over the indifference curves of the other (i.e. the union): an unmanageable bargaining externality.

Criticism 3: 'the model contains no explanation of why strikes occur'

The short response is that, indeed, the model of

[Section 2.3](#)

does not. Note, however, that this is not dissimilar to the economic literature on wage and employment target setting. For instance, McDonald and Solow (1981), Oswald (1983) and all other sources mentioned herein adopt a generalised Nash bargaining solution which in fact rules strikes out axiomatically! Though this was always recognised as an unsatisfactory feature, it was seen as a means of short-circuiting the well-known paradox of conflict theory (see Varoufakis, 1991,

[Chapters 5](#)

and

[6](#)

); namely that devising the perfect model of predicting conflict will, under rational expectations, abolish it!

In contrast to the mainstream neoclassical literature, the preceding analysis in fact

does offer a theory of why strikes occur – recall the whole of

Section 2.2

in which I adapted the approaches of Hayes (1984) and Kennan (1987); i.e. portraying strikes as devices for equilibrating information across the two sides, usually over the degree of expected worker mobilisation (as opposed to the state of product demand). Of course, this perspective – as we shall see in the next chapter – is utterly limited. I shall be arguing in

Chapter 5

that it is far less unpalatable to think of bargaining impasse as a by-product of evolving conventions for distributing a firm's surplus (rather than as the result of some instrumentally rational optimisation exercise). As we shall see, it is not impossible to model the feedback between distributional conventions and the conventions regulating relations amongst workers (even though some of neoclassical economics' holy cows need to be 'sacrificed' in the process).

2.3.8 Any connection with the reality of industrial disputes? You bet!

During a drawn out industrial dispute in the 1970s Cesare Romiti, head of Italian car giant FIAT, joined the picket line *incognito* to gauge the workers' mood. What he saw encouraged him to be intransigent in negotiations.

43

His judgment was vindicated when, some time later, the union's spirit was crushed by a white-collar march through Turin in favour of an end to the dispute. It only seems natural to suggest that economic models of trades union targets ought to pay as much attention to the *mobilisation rate* as Romiti did.

44

In recent years some neoclassical economists (a tiny minority, mind you) have recognised that the politics within a trades union matters, because it renders the conventional two-person bargaining paradigm incomplete. Oswald (1994), for example, utilised a median-voter scheme to introduce union politics into the debate. However, it seems doubtful whether a voting model of unionists' preference formation is the theory's main missing ingredient. Indeed, workers have continual influence over targets far more crucial than the occasional vote. They vote with their feet and in so doing they give, or deny, union targets relevance. Thus, regardless of *how* a trades union's collective preferences have come into being, there is no guarantee that labour's potential for collective action, and thus union bargaining power, will intensify the better a union's target reflects worker preferences.

2.4 Epilogue: the preceding analysis in the context of the meta-axioms' dance

As I explained in the introductory section (

Section 2.2.1

), I began working on the preceding models when, early in my career, I was struck by a remarkable 'absence' in the neoclassical literature on wage and employment determination in labour markets in which trades unions represent workers. What was missing was a recognition that collective action does *not* automatically spring out of a collective interest; something that neoclassical economists were only too keen to acknowledge in the context of all sorts of other analyses (e.g. their models of public good provision, of environmental degradation, of the fragility of cartels etc.).

Eager, as most young academics are, to 'plug this hole' in the literature, I set out to model wage and employment bargaining when worker mobilisation was neither automatic nor exogenous to the bargaining process, to the wage and employment targets pursued by the trades union, and to the social conventions binding workers together and, thus, enabling them to overcome free rider tendencies that would

otherwise wreck the very prospects of collective action which render the union movement relevant in the first place.

As the reader will have noticed in the preceding pages, my research programme bore fruit. Several papers were published on the subject and, whenever I presented these models in conferences, departmental seminars etc., the reaction from mainstream colleagues was positive. Except, my work ended up having precisely **zero** impact on the literature! The reason? It is very simple really. Neoclassical economists want one thing from models of wage and employment determination: a simple, well-defined function linking a host of 'purely economic' variables (such as marginal products of labour and capital, product prices, discount rates etc.) to the wage and employment combinations facing a firm. They need this in order to plug it into the profit function of firms and, from then on, to derive some reduced form that is potentially empirically testable. Alas, my analysis cannot produce this, for the simple reason that, as we have seen (recall, for example,

[Figures 2.7](#)

,
[2.8](#)

and

[2.9](#)

), the moment worker solidarity and mobilisation is subject to the social conventions that bind workers together, no such *one-to-one* and *onto* mathematical relationship can exist. Indeterminacy becomes the order of the day.

Indeed, not only does the introduction of the possibility of imperfect labour unity give rise to indeterminacy but it generates the most virulently radical form of indeterminacy as well. Neoclassical economists suspected, since at least the beginning of the twentieth century, that bargaining is, generally, an indeterminate process. But they thought that a trades union's targets *are* determinate, reflecting worker wage preferences and readiness to suffer striking costs to achieve them. However, my research above showed that *even these bargaining targets (let alone the outcomes of bargaining) are indeterminate!*

What should honesty compel economists to do once exposed to this analysis? Naturally, to confess that our models cannot, even theoretically, pin down the combination of wages and employment levels that will result from collective bargaining. That, when trades unions are involved, even to some extent, in labour markets, there can exist no mathematically determinate model that delineates the wage-employment schedule facing firms. Alas, such a recognition, even though uniquely consistent with 'scientific rigour', would mean the end of all econometric models that aspire to include an equation (or more) accounting for wages and employment. And that is not something that the 'profession' was ever going to accept lying down!

So, what did the profession do? Did it find some other theoretical fix? No, of course not (for no theoretical fix is possible). What economists have been doing ever since is to ignore the problem. To continue, as they did before the analysis above became available, to assume that wages and employment will fall either on the labour demand curve or on the contract curve, and to carry out their

econometrics *as if* trades unions are either totally absent or perfectly capable of mobilising each and every worker!

Summing up, when I started working on the models above, back in the 1980s, neoclassical economists had already asked many of the questions posed by this chapter. They issued the challenge represented by arrow **c** in the *dance of the meta-axioms* diagram in

[Chapter 1](#)

. Of course, very soon they retreated (arrow **r**), once they realised that such a move

unleashed irrepressible indeterminacy that threatened to wreck both the essence of their textbook analysis and, more importantly, their econometric models.

For my part, I took it upon myself to take on the challenge (arrow **c**), even though it led me straight onto the *Wall of Indeterminacy* of the *dance of the meta-axioms* diagram in

Chapter 1

. While I was happy simply to acknowledge this indeterminacy, and leave matters there, the profession was not. Thus, my analysis (despite some very polite noises from the profession) was condemned to remain off the agenda of 'serious economists' – the victim of an inexorable urge toward instant retreat from a worthy and logical challenge (arrow **r**).

VERDICT: Some of the cleverer neoclassical economists recognised the challenge of imperfect labour unity. However, pretty soon they chose to ignore it; to perform (what I described in

Chapter 1

as) the **1→2→1 quickstep**. And, when confronted with the models in this chapter, which were the result of my taking head on the 'challenge', the profession opted for an ignominious retreat from their findings, executing what I described (again in

Chapter 1

) as the **... 1→2→3→1 move**.

Notes

1

Based on Varoufakis (1989).

2

For example, see Kennan's (1987) extensive survey of the strike literature in which there is not a single reference to a theory incorporating imperfect worker mobilisation.

3

An imperfect but stable coalition will be more likely the more responsive the employer's last offer is to changes in σ relative to ρ , the lower the worth of psychological factors relative to monetary ones (i.e. γ) and the higher the anticipated strike length.

4

See Hayes (1984) for proof that a schedule of this form represents an equilibrium strategy for the trades union. The role of this schedule is to offer the firm incentive compatibility under circumstances of asymmetrically distributed information regarding the level of product demand.

5

The $\xi\lambda$ term contains the trades union's expectations on how the coalition will fare during the strike. In the pre-strike period workers communicate to them the value of $\bar{\gamma}$ which is used as a basis for computing the concession schedule (CS). The direction of change in σ will thus depend on whether γ' exceeds $\bar{\gamma}$ or not. With $\hat{\lambda}$ being the estimate of the speed of adjustment, $\xi\lambda$ A offers a complete description of the trades union's view of its future cohesion, solidarity and mobilization.

6

Proof of (

2.13

): $E[p(\sigma)] = \int \sigma p(\sigma) d\sigma = \int \sigma \int \pi(\lambda) f(\sigma | \lambda) d\lambda d\sigma = \int \pi(\lambda) \int \sigma f(\sigma | \lambda) d\lambda d\sigma$. Hence, $E(\lambda) = \int \lambda \pi(\lambda) d\lambda$. $\lambda \pi(\lambda) = \int \lambda f(\sigma | \lambda) d\sigma d\lambda = \int \lambda \pi(\lambda | \sigma) p(\sigma) d\sigma d\lambda = \int E(\lambda | \sigma) p(\sigma) d\sigma$ i.e. the expectation of the prior of λ is equal to the expectation of the posterior averaged by the Bayesian predictive distribution. Alternatively, $\hat{\lambda} = \int \int \lambda f(\sigma | \lambda) \pi(\lambda) d\sigma d\lambda$. QED

7

If the firm did not possess perfect information regarding the shape and location of function ρ , then an adaptive learning

rule is as good as any: $\frac{d\hat{\lambda}}{ds} = \beta(\lambda - \hat{\lambda})$.

8

Kreps and Wilson (1982) were the first to make this point in the game theoretical literature. They showed that in finitely repeated prisoners' dilemmas, backward induction is prevented from wrecking the chances of cooperation if a modicum of uncertainty is injected in players' minds, leading them to believe that a non trivial probability exists that their opponent prefers mutual cooperation from defecting against a cooperator.

9

Thom's theorem shows that if adjustment is not instantaneous in a model governed by smooth functions and containing two parameters, there is essentially only one possible type of geometrical structure (see Poston and Stewart, 1978, for the full exposition): the cusp of

Figure 2.7

, where the two parameters in question are the anticipated average profitability over the whole dispute (m) and a measure of the gradient of the trades union's CS (η). For other economic applications of the cusp, see Harris (1978) and Dogson (1982).

10

Note that although

Figure 2.7

is topologically the most complicated structure, there is always the possibility of multiple cusps. If, for example, we used the firm's discounting rate (r) as an additional control variable, it would be possible to generate a second cusp for high values of r . However, this would not be particularly interesting since it is not very likely that r will change during the dispute.

11

Based on Varoufakis (1990) as well as a series of working papers on the same subject that followed over the next eight years.

12

Originally Dunlop (1944) and later Nickell and Andrews (1983).

13

See McDonald and Solow (1981).

14

For example, Oswald (1984). See also Blair and Crawford (1984).

15

See Svejnar (1986) and Manning (1987).

16

See Brown and Ashenfelter (1986), Card (1986), McCurdy and Pencavel (1986).

17

See Doiron (1992).

18

The aggregation of members' preferences over (w , N) is not discussed here. Some political process is assumed to have given rise to function $W(\cdot)$; perhaps through the election of shop stewards and representatives.

19

Of course the strike will not happen when the willingness of a proportion σ^* of the workforce to walk out for s periods is common knowledge – in which case rational negotiators will settle without a dispute on a wage/employment level that reflects the (s , σ^*) threat. However if the mobilisation rate is not common knowledge at the outset, a strike can be thought of as the mechanism that equilibrates beliefs. See section 2.

20

Implicit in this model is that union leaders and union members are identically informed. Thus inequality (

2.21

) is a pre-condition for each worker to compute their individual decision rule in (

2.16

) on the basis of the announced (w , N) union target. I return to this in the next section.

21

The precise functional form of the slope of the firm's isoresistance curves is:

$$\frac{ds}{d\sigma^e} \Big|_{d\Pi=0} = - \frac{1 - \exp(-\delta s)}{\delta \exp(-\delta s)} N_0 \times \frac{w_0 - p \frac{\partial Q}{\partial N_{|N=(1-\sigma^e) \bullet N_0}}}{p[Q(N_0) - Q((1-\sigma^e)N_0)] + C_s - \sigma^e w_0 N_0}$$

From this we can see that, as long as (i) strike-breakers' marginal revenue product exceeds their wage (which is presumed to be the same as the pre-negotiations wage) and (ii) the value of the lost output per strike period (including the intangible losses C_s) exceeds the wage bill savings per strike period, the firm's isoresistance curves will be downward sloping. In the case where (i) does not hold, employers have no reason to employ strike-breakers during a dispute and this is reflected in either flat or upwards sloping isoresistance curves (at least for values of σ above a certain threshold) on the (s , σ^e) plane.

A sufficient condition for concavity is that, the overall losses to the firm per strike period per (pre-strike) employee exceeds

$$\frac{(w_0 - MRP_{SB})^2}{\frac{\partial MRP_{SB}}{\partial \sigma^e}}$$

i.e. the ratio of the square of the difference between the wage and the marginal revenue product of strike-breakers over the rate of increase of the strike-breakers' marginal revenue product with every worker who joins the strike. In case the above does not hold, concavity will still prevail as long as:

$$(1 - e^{-\delta s}) \cdot \frac{(\text{Net Strike losses}) \cdot \frac{\partial \text{MRP}_{\text{SB}}}{\partial \sigma^e} - (w_0 - \text{MRP}_{\text{SB}})^2 \cdot N_0}{\text{Net Strike Losses}} \\ + \frac{(w_0 - \text{MRP}_{\text{SB}})}{\text{Net Strike losses}} \frac{\delta e^{-\delta s} + (1 - e^{-\delta s}) \frac{\partial \text{MRP}_{\text{SB}}}{\partial \sigma^e}}{\delta e^{-\delta s}} > 0$$

22

For greater clarity I assume that the union is large enough so that each worker does not consider her decision alone to have the capacity to sway the balance of power between the union and the firm. This assumption could be relaxed by, for example, allowing each worker to believe that her decision to strike would increase the negotiated wage (or level of employment) by a magnitude whose effect on her inter-temporal income is minuscule when compared with her wage losses from striking.

23

For simplicity I assume that the announced wage targets are sincere and do not reflect either an ambit claim or exaggeration or indeed irrational wishful thinking. Alternatively, the reader can interpret the union's targets (w , N) which feature in the workers' maximands as their own rational expectations of what union leaders aim at given the publicised (inflated) objectives.

24

Note that, however close the individual's preferences may be to (or distant from) the union's targets, the act of joining a strike is a dominated strategy unless crossing picket lines comes at a (non-pecuniary) cost; i.e. unless $\rho > 0$. The significant implication of this point is that a worker's preferences viz. the long term general level of wages and job security do not affect one's decision to strike or not. The latter is affected only by short-term factors which are decided by their decision to strike; e.g. the possibility of being victimised by the employer if one strikes, the lost wages during the strike, and the lost reputation amongst one's colleagues. Another way of rationalising the individual's intention to strike would be in terms of a tit-for-tat solution of the free-rider problem. However, in view of the multiplicity of equilibria in such models (i.e. the Folk theorem), the rest of the model would be intractable. Moreover, the idea that workers care in important, non-pecuniary, ways about their stance during a dispute seems more plausible than a trigger-strategy explanation.

25

For example, if only a few workers break the strike, the firm may reward them with a wage higher than w_0 while the union's capacity to pay benefits to strikers will be circumscribed. In the opposite case, where σ is rather low, the $(w_0 - b)$ differential could be smaller.

26

i.e. assuming that, if γ^{\min} is the lowest γ value, then $\gamma^{\min} \rho(\sigma = 1) > z(\sigma = 1)$.

27

Note that σ^* is also an equilibrium mobilisation rate. However, under (

2.17

) it is unstable.

28

Where $N(\cdot)$ is the truncated normal distribution, η_m and η_σ are the mean and standard deviation of the η s, and the truncation is such that $\eta_i \in [0, 1/\sigma^*]$.

29

Note that this model assigns an important role to union leaders which has been neglected in the literature: with good rhetorical and political skills they can instil greater optimism in their members' hearts (i.e. reduce η_m) and thus increase the probability of successful mobilisation *ceteris paribus*.

30

With disunity equilibria, mean optimism plays no role in determining mean mobilisation since the interior equilibrium σ^* is an attractor independently of the value of η . In such cases, the only role optimism can play is to speed up convergence to the equilibrium level of mobilisation.

31

Indeed unions may find that their bargaining power is greater when they threaten managers with a shorter strike duration provided the expected mobilisation rate is higher.

32

As σ^* is the mobilisation level that solves $\gamma_m \rho(\sigma) = z(\sigma, s)$, any increase in s boosts $z(c, s)$ and in the case of **unity equilibria** increases σ^* thus reducing σ^e – see equation (7). Under **disunity equilibria** increases in s force σ^* to decline and therefore, again, lead to a reduction in σ^e – see equation (2.19)

33

This is so even under perfect information and risk neutrality in which case the union has no reason to fear living life on the edge; that is, imposing (provided it can do so) a bargain which brings the firm close to closure.

34

That is, if the union leadership were to foreshadow either a shorter or a longer potential dispute, the resulting rate of mobilisation would be insufficient to scare the employer into accepting the chosen target.

35

Note that this equilibrium is of the **disunity** kind – see

Section 2.3.4

36

Why would the union's power ever increase as its target shifts from X to E or to D? Because some times a wage (or employment) target below what is considered amongst the workers to be 'fair' or 'right' reduces the 'pain' of crossing picket lines. Technically speaking, when this is so a shift from X to E or D may cause the *solidarity function*, i.e. $p(\cdot)$, to rise faster than the *loss function*. This ceases to be so, by definition, when the target reaches the *boundary*.

37

Perhaps the question ought to be the opposite: Given that the union's target will belong to the LCC only if the tangency point between a firm's isoprofits and a union indifference curve happens also to be a tangency point between the latter and the union's *boundary*, why has the literature so readily accepted that efficient union targets will lie on the LCC? The answer, of course, is that the literature treats ψ as a single point subset of the LCC (since worker mobilisation is assumed to be perfect or, at least, exogenous).

38

Alternatively, in the middle of a recession our worker may privately prefer the \$350 wage target but find it harder (for purely non-pecuniary reasons) to cross picket lines when the union campaigns for \$300 as well as for 10% extra employment which will benefit unemployed colleagues (or colleagues who are facing the axe). In this case, provided this is typical of most working members, the union = *s optimal target* may lie on the right of the LCC.

39

The reader may object that the reason why the LCC no longer contains the union *optimal target* is that I have introduced socially determined wage preferences into the worker's utility without allowing the LCC to reflect these new preferences. This is not so. Our individual worker does not care about the socially determined view on what constitutes a fair, or proper, wage/employment level. She only cares about being seen (even by her own eyes) to be breaking a strike which has community support. Thus the 'fair' wage and/or employment levels do not affect private wage and/or employment preferences in any defensible manner. Indeed if the expected level of mobilisation is low (or zero), our typical worker will not think about them twice.

40

Interestingly, the decision of workers to support strikes disproportionately to their private wage and employment preferences does not even need to be an act of altruism: our workers may simply not want to be *seen* to be crossing the picket line under these circumstances. Analytically, such thoughts enlarge the gap between aggregate preferences over (w, N) and the degree of mobilisation the leadership can anticipate for that same target.

41

Note that this argument applies equally in the case where the LCC and the LDC coincide – as in Oswald (1994).

42

With an infinite horizon of future negotiations ahead, or even with the thought that an employer who detects a union leadership lacking credibility amongst its members will immediately seek to reverse its gains, it is possible to show that honesty on the part of leaders is a perfect equilibrium strategy. Suppose however that union leaders are elected every few years for a fixed term of n negotiations. During the last negotiation (n), leaders have no reason to abide by their target announcements and, therefore, workers will expect them to re negotiate towards the LCC. Backward induction recommends that, in all n negotiations, workers will assume that contracts will converge on the LCC and will never believe their leaders protestations that they intend to stick to pronounced targets off the LCC [e.g. D in

Figure 2.8(d)

]. However, if workers entertain even a small subjective probability that leaders take pride in being seen to be honest, then it can be shown (e.g. see Kreps and Wilson, 1982) that there will be at least $m < n$ negotiations during which workers will believe their leaders' target. Moreover, even leaders who do not enjoy being truthful will stick to their targets with a view to breaking their promises towards the end of their tenure. As the end of the horizon approaches, the probability that leaders will renegotiate (and thus return the union to the LCC) increases. The efficient bargaining story will thus gain poignancy towards the end of a leader's horizon (provided leaders are indeed impervious to their reputation for honesty) and disappear again when the new leader is installed. Perhaps another determinant of wages and employment under trades unions has been revealed: the frequency of union elections and the elected leaders' credibility.

43

In an interview for *The European* newspaper (5–11 December 1996) he was quoted thus: 'It was a bit risky, and basically a bit stupid. But when I got there I saw that the pickets were made up of the professional agitators, not the FIAT workers. I went home thinking that things were going to go our way.'

44

During the 1984 UK miners' strike, newspapers, radio and television reported daily the precise *mobilisation rate*. It seems that reporters recognised that the outcome would hinge on the dynamic path of this rate.

3 Rational conflict

On the impossibility of a determinate theory of costly disagreement

3.1 Prologue

3.1.1 Background briefing

The previous chapter investigated the formation of bargaining targets in the context of disputes between labour and capital, between trades unions and employers. It turned on the idea that, in forming targets and determining the relative bargaining power of each side, worker mobilisation and solidarity was crucial. What it did not do was to offer a model, an analytical explanation, of why actual conflict occurs. Come to think of it, it is one thing to model the bargaining power of a union or of the firm's management and quite another to explain why the two sides will allow mutually damaging conflict to occur (instead of settling their differences *sans* conflict). Could a formal theory of conflict be assembled out of the concepts and methods of mainstream economics? Could game theory, the highest form of neoclassicism, be the source of such a theory?

Rational Conflict was my first single-authored book (Varoufakis, 1991). The idea of writing it came to me when I encountered a delicious antinomy buried deeply in the foundations of game and bargaining theory:

The paradox of rational conflict

Suppose there exists a splendid theory of conflict, say *T*. Theory *T* can be used to work out the optimal bargaining strategy of each party to any negotiation at every point in time. Consequently, *T* can also yield estimates of (a) the final agreement *A* that will be reached and (b) the length of time *t* that it will take to reach it. Moreover, if bargainers are rational, each one of them ought to have (mental) access to theory *T*, which means that each has a similarly accurate, and common, estimate of both *A* and *t*. But if they share a common estimate of the final agreement, *A*, and delays in reaching it (period *t*) are costly (e.g. a strike), then rational bargainers ought to agree instantly on *A*. In short, if a uniquely rational theory of conflict, *T*, exists, then rational bargainers must never allow their negotiations to lead to costly disagreement (i.e. $t = 0$). The meaning of this conclusion is that either there can be no such thing as a uniquely good theory of conflict or that conflict is the result of irrationality.

The gist of the above paradox is that, seemingly, if we could develop a brilliant theory of conflict, then the possibility of rational conflict (that is, of conflict

between rational agents) would, necessarily, wither as rational antagonists would have no reason to go through the motions of 'fighting.'

This paradox struck me as an excellent opportunity to cast a critical gaze on the foundations of game theory in particular and neoclassical theory in general. What I discovered, when I looked at these foundations carefully, was particularly unappetising. Thus began a twenty-year engagement with some of the basic notions of game theory, at a time when game theory was becoming all the rage and was leading neoclassical economists to make huge claims about their discipline (e.g. the claim that game theory can unify all the social sciences, rendering neoclassical economics the official queen of social theory).

The purpose of this chapter is to explore the extent to which a logically defensible neoclassical theory of conflict (based on game theory) may be impossible. In short, the paradox of rational conflict cannot be overcome without taking liberties with the rules of logic. And, since conflict between perfectly rational agents has been known to occur, this leaves a great big hole at the centre of any theory of society which remains exclusively neoclassical.

3.1.2 The rest of this chapter

The next section asks the question: Why would rational people allow their differences to spill into costly conflict? Can game theory account for such instances of 'rational conflict'? On what basis could we bypass the paradox of rational conflict, as stated above.

Section 3.3

looks at so-called non-cooperative games and shows that conflict is perfectly possible when bargainers have no capacity to reach a binding agreement by means of well-run negotiations. But then, of course, the question becomes: Why would they not engage in such negotiations, and why would they not subject themselves to a third party's or institution's authority if doing so would make binding agreements possible?

Section 3.4

asks whether the availability of binding agreements would, in fact, eradicate conflict.

Section 3.5

argues that the whole analysis of rational conflict, if we demand that it is conclusive (and that the models are 'solved' or 'closed') is predicated upon assumptions which render conflict... impossible. Section 3.6 concludes and links this literature to our *dance of the meta-axioms*.

3.2 Conflict defined and game theory's potential introduced

Conflict has traditionally caught the imagination of scholars who felt the need to delve into its causes long before economists puzzled over relatively innocuous problems such as inflation and unemployment. Nevertheless, the merits of economic analysis are often presented in terms that the non-economist student will relate to. Economics, to be precise, is canvassed as the study of how agents come to an automatic settlement of antagonistic interests caused by scarcity. It is when agents relentlessly strive toward their personal interest, with little or no concern for the social effect of their actions, that the public good is best served.

The economist construes the ideal social world as one surfacing because, rather than in spite, of the pervasiveness of the individual's belligerence. The key to the prevention of the transformation of these 'natural' tendencies into conflict is, of course, the work of Reason. Provided the institutions of the market are in place, individuals are supposed to harness their instincts and, in doing so, forge a splendid resolution free of wasteful activity. Paradoxically, conflict is at once the guarantor of optimal social outcomes and the cause of its own demise. Equilibrium and stability are, therefore, the by-products of a social order founded on the disposition to fight for one's self. Philosophers understand the above as the alleged supra-intentional work of Reason, and recognise that economics claims to have done their work for them.

Indeed, one is excused to think that, as far as economics goes, an explanation of conflict is not urgent; conflict has been designed out of the system. Industrial strikes, hostile takeovers, oligopolistic wars between firms, trade and currency wars between nations are all, supposedly, minor nuisances that do not threaten the miracle of the market. Just as the odd meteorite does not give cause for changing the calculations of the length of Mars' year, so too the waste of resources inherent in the advertising war between Coca-Cola and Pepsi-Cola, or Apple and Samsung, offer insufficient reason for questioning the market's ability to eradicate conflict.

Be that as it may, if economics is to dominate the social sciences, as is the economists' wont, surely economists must have something meaningful to say about the occasional emergence of strife; after all, astronomers do not remain mute about minor celestial phenomena. It is, however, the contention of this chapter that, unlike astronomers, economists have little to say about conflict that (a) makes sense and (b)

does not undermine their own research agenda.

Given the large dose of interdependent behaviour that is required by any analysis of conflict, game theory was bound to be employed in order to elucidate conflict on behalf of the economics profession. For game theory is the highest form of neoclassical economics' analytical method. It begins with standard marginalist models of choice but then allows the outcomes of Jill's choice to depend on Jack's choices too. And, since Jill is assumed to be rational she knows that her own choice must depend on her beliefs about what Jack is up to. And she knows that Jack knows that, thus ensuring that Jill will choose her actions in a manner that reflects what she believes that he believes that she will be choosing. And so on and so forth. The question then is: What can game theory, which valiantly takes on this Gordian Knot of beliefs, do to account for conflict?

Before answering this question, it is imperative that we define conflict. When an agent acts in a manner that destroys part of a valuable resource with a view to enhancing personal gain, we can safely claim to have observed an instance of conflict. Wars and strikes fall under this category. However, there is another less obvious category. When agents fail to reach an agreement that would have given rise to a net increase in their joint stock of value or wealth, conflict is also in the air. However, this definition may have licensed too much. One could protest that the second category classifies every process leading to non-Pareto outcomes as being indicative of conflict. In fact, if we accept this definition, the 'small'

question on conflict begins to touch upon some very thorny issues. For every dynamic disequilibrium process entails some Pareto loss, at least before equilibrium is re-established. In terms of my definition of conflict, there is a relevant dimension in economic theory going back to Adam Smith, David Ricardo and Karl Marx. Classical economists, for instance, saw the equilibrium level of prices ('natural' for Smith, 'productive' for Marx) as a function of the type of adjustment. In more recent terminology, the out-of-equilibrium behaviour of the system shapes the actual equilibrium. Consequently, an identification of conflict with welfare losses confers an abstract view of the market economy where conflict, though transient, plays a crucial role in determining the kind of harmony to which the choices of sovereign agents inexorably lead. Game theory is seen by many as an opportunity to provide a rationale to such quasi-functionalist speculation.

Returning to the definition of conflict, it seems inevitable that we must consider as an instance of conflict any combination of choices resulting in deadweight losses. By deadweight losses I mean that the outcome is worse for the individual agents involved compared to what it could have been had they chosen differently from within their set of feasible choices. Whether this is possible when the agents are rational is equivalent to asking whether there can be such a thing as 'rational conflict'.

The essence of game theoretical reasoning is that agents intelligently assess the effects of their choices on the behaviour of their opponents before acting. This is a commendable departure from traditional myopic reaction functions but, unfortunately, it is not enough. Progress along the game theoretical path requires that agents *replicate* each other's thoughts. The presumption that rationality is in place and commonly known allows game theory to use the notion of equilibrium in order to cut the Gordian knot of interdependent behaviour. Unfortunately, the same presumption undermines the relevance of the theory.

Although equilibrium conflict may sound like a contradiction in terms, noncooperative game theory gives it its head. If I think (a) that you can replicate my thoughts, (b) that whatever I do you are better off shunning peace in favour of violence, and (c) that if you choose violence I am better off doing the same, then violence is the equilibrium outcome. Nothing can prevent this instance of counter-finality from conducting war and driving a wedge between the Pareto and the Nash equilibrium outcomes. However,

some equilibrium outcomes are more paradoxical than others.

Confronted with the threat of an attack, our agent is alarmed. If the potential aggressor is kind enough to announce in advance that the assault will occur either today or tomorrow, but that it will only take place provided that it is not anticipated with certainty on the day when it will occur, our agent is reassured. She has reasoned that, if the attack does not eventuate today, then it cannot take place tomorrow either because it will be anticipated, thus violating the condition for the attack. Moreover, given that it cannot occur tomorrow, today is ruled out too for the same reason – the paradox of backward induction, as it is known in the literature. According to game theory's equilibrium approach, our agent is right to feel safe. Ironically, now that the agent is safe in the thought that no assault will

occur, the assault can now take place without bending the rules of engagement. But if this is so, and it is, then the agent must start worrying again that she is about to be attacked. When should she expect this attack? Tomorrow? No, since (by the logic outlined above) no attack tomorrow is logically justified. In which case there will be no attack today either. Should she feel safe again? No, because if she does then she will be attacked. But when should she expect the attack? And so on and so forth, the agent's logic going around an endless circle resembling standard paradoxes like 'I am a Cretan and all Cretans are liars'.

Moving beyond non-cooperative game theory, cooperative or bargaining theory comes into play when agents have the capacity not only to negotiate but also to reach binding agreement. As one might expect, this form of game theory offers even fewer opportunities for a theory of conflict. Indeed, as I shall be arguing below, ultimately bargaining theory (also known as cooperative game theory) can never explain conflict between rational agents. Not even if they are asymmetrically informed. The deeper reason for this is captured nicely by the paradox of rational conflict, as presented in the previous section. Let us look more closely at both non-cooperative and bargaining theories' capacity to throw light on why rational people fight. The next section concentrates on non-cooperative game theory while

[Section 3.4](#)

examines bargaining theory's limits.

3.3 Equilibrium conflict I – non-cooperative game theory

3.3.1 Nash's equilibrium concept

The neoclassical economic method consists of defining agents' objectives and constraints and, then, proceeds to discover the set of their choices that are optimal or, equivalently, that constitute an equilibrium set. Can conflict result when agents converge on such equilibrium strategies? Before answering the question, let us familiarise ourselves with John Nash's conception of a game's 'solution': what he referred to as an *equilibrium*.

Suppose that each player must choose an 'action' or 'strategy' or 'move' from a (finite) set of such choices (henceforth I shall refer to these as 'strategies'). Suppose further that rational thought can lead each one of them (along with us, the theorists) to a unique conclusion as to which strategy it is in her interest to choose. In this case, it is *as if* the players' thought processes have converged to an equilibrium, just as surely as a rock tumbling down a hill eventually reaches an equilibrium (a 'state of rest') on the hill's foot. Thus, a game's equilibrium is conceptualised as a set of strategies, one per player, such that the more rationally each player thinks of her 'situation' the more she tends to converge on the specific strategy in that set.

To give an example, consider the following simple N -person game known as the *Race to Zero*: N players are asked to write on a piece of paper (in isolation from one another) a real number between 0 and 100 (inclusive). The player whose chosen number is nearest the *maximum choice* among all players *divided by 2* wins \$1

million times her choice of number. (Joint winners divide the spoils.) Is there a 'solution' to this game? Is there an equilibrium toward which the players'

choices will tend the more rationally they think? What number should one write down?

Nash suggests that rational players would immediately decide that it makes no sense to choose a number in excess of 50; to think that: 'Since the largest number that can be chosen is 100, and I win if my choice is nearest to that maximum choice divided by 2, I should never choose a number above 50.' However, this thought immediately begets another, infinitely longer, thought:

'If I am clever enough to work this out, then the rest will also work this out too. Therefore, none will select a number greater than 50, in which case I must not choose any number above 25. But if this is so, will the others not know this to be so too? And if they do, will they not restrict their choices to a maximum of 25? Then I must not go beyond 12.5.'

And so on. Asymptotically, one's optimal choice of number tends to zero just as surely as the proverbial rock rolls down a hill until, asymptotically, it hits rock-bottom: 'Choose zero' is, therefore, the game's equilibrium.

To sum up, in this case of strategic uncertainty, one's estimation of how others think is crucial. Had one's opponents been mindless machines, or monkeys, the only certainty is that one ought not select a number above 50. But, when playing against other rational players, and knowing it, a logical chain reaction leads each player to the choice of zero. Equal winners of exactly nothing! The impetus to this ruthless outcome is none other than infinite order common belief in instrumental rationality (**CBIR** hereafter): As long as one believes that all others believe that one believes that all others believe ... [*ad infinitum*] that everyone is instrumentally rational, they all choose zero.

Nash's brazen theoretical move, which allowed him to get to this unique equilibrium, was simple: He *rejected all beliefs which, if held, would lead to behaviour that would have falsified these beliefs*. Put differently, he *admitted only beliefs which will be confirmed by the strategies which they recommend*. Put differently again, Nash *assumed that rational players, who recognise that their competitors are also rational, will never expect them to hold false beliefs*. In the above game, it is easy to see that if one follows Nash's lead and discards all beliefs which would be contradicted by the group's choices, there is only one left:

² the belief that each will select zero. When all players believe this, each chooses zero and the *Nash equilibrium* materialises.

³ The elimination of all 'false' beliefs does not only solve the *Race to Zero* (by eliminating all strategies per player except one); it also helps illuminate Adam Smith's argument that the *invisible hand* surreptitiously eliminates the merchants' profits (just as it had led the players in the *Race to Zero* to actions that eliminated their winnings), thus delivering the lowest possible prices for consumers. However, at the very same time, the same Nash equilibrium concept can explain why people, left to their own devices, may fall into the Leviathan trap; i.e. be allowed to be trapped into a mutually damaging conflict.

3.3.2 A conflict game

Consider the following game in which the two agents, R and C again, have a choice between three strategies: non-violent, medium intensity conflict and high intensity conflict. Clearly, the peaceful outcome is preferable, for both of them, to either conflict outcome (note that their payoffs are higher compared to those in the other two symmetric outcomes: (4,4) as opposed to (3,3) and (2,0) respectively). Will peace prevail?

There is no doubt that R and C are better off in peace than in war, as their individual payoffs are higher in the top left hand side of the matrix's diagonal than in the other two cells. And yet, Nash's equilibrium logic suggests that peace is doomed. Indeed, following Nash's method (see above), it is easy to show that the only outcome here that constitutes an equilibrium is that in which both sides opt for medium intensity conflict.

To see this, observe that the top left hand side outcome (3,3) is the only one which is not supported by false beliefs (i.e. predictions) on either side: Take the non-violent outcome, to begin with (4,4). To occur, the party choosing from the rows must choose non-violence. Would she do this if she expected non-violence from her opponent? Sure she would, as non-violence is the row player's best reply to the column player's non-violence. But, this is not so for the column player whose best response to the row player's non-violence is 'high-intensity conflict' (observe that the column player would then receive payoff 0,5). By a process of examining, in this manner, the correspondence of actions and beliefs each cell in this game matrix we come to the conclusion that there is only one Nash equilibrium: Medium-intensity conflict. [Check: both players' best reply to the belief that the other will choose medium-intensity conflict is to choose medium-intensity conflict themselves.]

The question is: Granted that Nash's equilibrium method provides us with a determinate 'solution', i.e. prediction, does it make sense? Or, more precisely, does it own a monopoly of the truth regarding how R and C may behave reasonably? My answer is that, while the Nash equilibrium is the only equilibrium of this game it is not the only rational outcome. Indeed, R and C may opt for peace, quite rationally, even if peace is an out-of-equilibrium outcome. If I am right, a wedge will have been driven through the idea that a game's, or model's, rational solution must necessarily be an equilibrium of the game. Put differently, if I am right that rational peace is possible here, even though it is not in a Nash equilibrium, then the explanatory and predictive power of Nash's equilibrium concept withers to insignificance and radical indeterminacy takes over as any of the outcomes in

Table 3.1

becomes possible. Let us ask ourselves a series of simple questions: Could R choose the on-violent strategy R1, even though it is out of equilibrium, and have a solid rationale for doing so? The answer is affirmative and comes in the form of the following justification:

I will play R1 because I expect C to play C1. Why do I expect this? Because I do not think C expects me to play R1; indeed I think he expects that I will be playing R3 (rather than the R1 which I intend to play). You can ask me why I think that he will think that. Well, perhaps because he expects that I will mistakenly think that he is about to play C3, when in reality I expect him to play C1. Of course, if he knew that I was planning to play R1, he ought to play C3. But he does not know this and, for this reason, and given my expectations, R1 is the right choice for me. Of course, had he known I will play R1, I should not do so. It is my conjecture, however, that he expects me to play R3 thinking I expect him to play C3. The reality is that I expect him to play C1 and I plan to play R1.

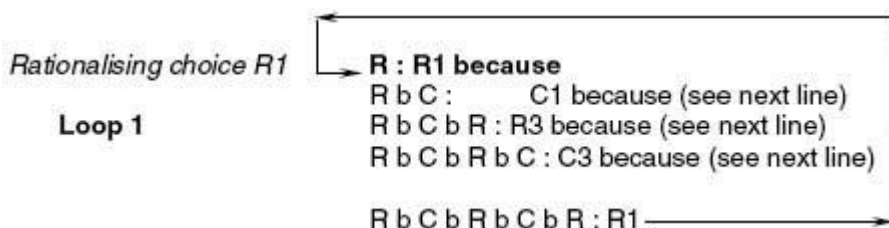
Table 3.1

Rachel and Charles choosing between war and peace

<i>Rachel's (row) and Charles' (column) strategy choices</i>	<i>C1: Non-violence</i>	<i>C2: Medium intensity conflict</i>	<i>C3: High intensity conflict</i>
R1: Non-violence	4, 4	2, 3	0, 5
R2: Medium intensity conflict	2, 2	3, 3	1, 2
R3: High intensity conflict	3, 4	2, 0	2, 0

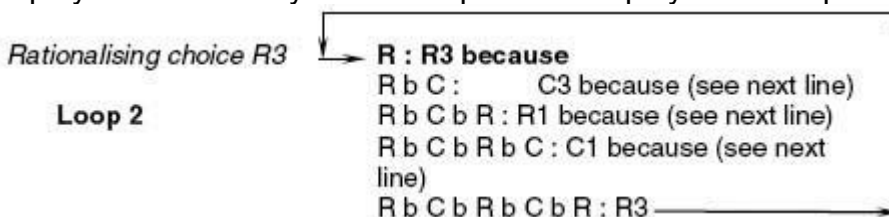
Note: The shadowed cell denotes a Nash equilibrium.

The above thought process can be summarised, using a simple shorthand (according to which 'b' stands for 'believes' and ':' for 'chooses' or 'will choose') as follows:



Next question: Could R have chosen the most belligerent strategy, R3, and still have a solid rationale for having done so? Indeed she could and it would take the form of the following argument, complete with its shorthand version that follows below:

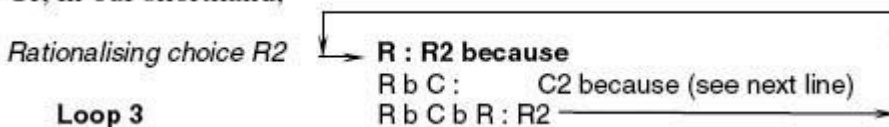
I will play R3 because I expect C to play C3. Why do I expect this? Because I do not think C expects me to play R3; indeed I think he expects that I will be playing R1 (rather than the R3 which I intend to play). You can ask me why I think that he will think that. Well, perhaps because he expects that I will mistakenly think that he is about to play C1, when in reality I expect him to play C3. Of course, if he knew that I was planning to play R3, he ought to play C1. But he does not know this and, for this reason, and given my expectations, R1 is the right choice for me. Of course, had he known I will play R3, I should not do so. It is my conjecture, however, that he expects me to play R1 thinking I expect him to play C1. The reality is that I expect him to play C3 and I plan to play R3.



Does any of the above mean that R will *not* be choosing R2, as predicted by Nash's equilibrium outcome? Of course not. She may very well play R2 on the basis of the following thought process:

I will play R2 because I believe that C will play C2. And why do I believe that C will play C2? Because he thinks that I will play R2, thinking that I expect him to play C2. And so on.

Or, in our shorthand,



As we just saw, all three strategies (R1, R2 and R3) can be fully rationalised (game theorists refer to them as *rationalisable*) because there are plausible beliefs that support each strategy. But if any of the available strategies can be rationalised, this means that the analysis cannot tell us anything about what a rational R will do in this game (and, similarly, C's choices are equally open). In short, the game is *indeterminate*.

3.3.3 A leap of faith

We have come to the crucial point: To tell a determinate story of conflict we need to argue that (commonly known) rationality helps us narrow down the range of potential outcomes, hopefully to one. In this game, intellectual honesty compels us to admit defeat in the hands of radical indeterminacy. And yet this is not something neoclassicists are prepared to do. Instead they are prepared to bend the rules of logic in order to get what they want: a determinate solution. How do they do this precisely?

In the case of non-cooperative games, like the one in

Table 3.1

, John Nash provided neoclassical economists with a theoretical trick that they use in order to pretend that the indeterminacy has been defeated. Nash notes that, among the set of rationalisable strategies (R1, R2 and R3 for R and C1, C2, and C3 for C), one strategy pair stands out: (R2, C2). Why does Nash think that R2 and C2 are particularly salient for R and C? Its unique appeal springs from an interesting feature: R2 and C2 are strategies supported by beliefs which will *not* be frustrated by the actual choice of R2 and C2.

To see this clearly, recall that strategy R1 is rationalised by R thinking (a) that C expects her to choose R3 and (b) that C will choose C1 in reply. There are two possibilities. Either R's predictions are confirmed, or they are not. Suppose they

are (and that R has indeed played R1). In that case, C's beliefs will be frustrated. We know this because the only way R's beliefs could be confirmed is if C has played C1. Why would he do this? The only rational belief that would make C play C1 is if C expected R to play R3. But R has frustrated that belief of C with his actual choice of R1. Alternatively, R's own beliefs will have been frustrated (when C chooses something other than C1). In either case, *the play of R1 will frustrate someone's beliefs and will only be played rationally by an R who is confident that the outcome will frustrate her opponent's beliefs, rather than her own.*

Of course, the same applies to R3 since, as we saw above, it also relies on a logical loop according to which R selects R3 on the strength of her belief that C will *not* expect her to play R3. But the same does *not* apply to R2. In fact, by choosing R2, R is telling the world that she is expecting C to make no mistake in predicting her strategy.

In other words, R2 is chosen rationally only when R has no reason to think that C will base his decision on a mistaken prediction of her choice. Similarly, C2 will be played when C has no reason to predict that R will be fooled. And when R and C choose R2 and C2, their actions will confirm their trust in one another's capacity to avoid erring.

Having said all that, why should we assume that rationality (even when commonly known) can forge an alignment of beliefs between the two players which gives the cause to believe that they cannot fool one another? The honest answer is: It cannot! But if neoclassical economists were to accept the truth of this simple statement, they would have no determinate theory of strategic choices in this, and other, conflict games. So, game and neoclassical economic theorists swallow their pride, set aside the rules of logic and assume that rational players will never allow themselves to imagine that they can fool one another; even when there is no logical reason to expect that they cannot.

Don't neoclassical economists recognise this? Of course they do. For instance, Kreps (1990) had this to say on the matter:

We may believe that each player has his own conception of how his opponents will act, and we may believe that each plays optimally with respect to his conception, but it is much more dubious to expect that in all cases those various conceptions and responses will be 'aligned' or nearly aligned in the sense of an equilibrium, each player anticipating that others will do what those others indeed plan to do.

What this leading game theorist is effectively admitting is that, just as anyone who has talked to good chess players (perhaps *the* masters of strategic thinking) will testify, rational persons pitted against equally rational opponents (whose rationality they respect) do not immediately assume that their opposition will never err in predicting their behaviour when their behaviour is inherently indeterminate (and thus impossible to predict). On the contrary, the point of good chess players is to *engender* such predictive errors!

Nevertheless, and this is the crucial point, despite such admissions by the top game theorists, game theory and mainstream economics continues to assume that games like that in

Table 3.1

have a determinate outcome (medium-intensity conflict, or R2, C2). Why? Because neoclassicists loathe one thing more than they do intellectual dishonesty: indeterminacy!

3.4 Equilibrium conflict II – bargaining or cooperative game theory

3.4.1 Does the availability of binding agreements eliminate conflict?

Taking stock on the question of what game theory can offer in terms of a theory of rational conflict, the previous section yields two important conclusions: First, there are many instances when, even when a single equilibrium exists, we still have no clue of what will happen – of whether agents will settle without a fight or not. Indeterminacy, in those cases, is all-consuming. Secondly, in games where there is a unique equilibrium, as in the game of

Table 3.2

(where outcome R2, C2 is the only rationalisable solution), the question emerges: What if players could reach a binding agreement that would allow them to settle without a dispute, e.g. to agree to a peaceful outcome in

Tables 3.1

or

3.2

and then divide between them the overall gains? Can game theory ever explain why they may fail to reach such agreement? Is a theory of conflict between rational bargainers possible? This section seeks an answer to this fascinating question.

Let us confine our attention to the many ‘games’ people play in which *binding agreements* are possible *prior to action* thus enabling players to reach decisions *jointly*, and by negotiation; as opposed to *competitively* (or, in the game theorists’ own language, *non-cooperatively*). For example, organisations usually converge on action-plans on the basis of collective deliberation and bargaining, and not merely through autonomous choices by isolated individuals (like those in the games discussed so far in this chapter). States too possess means of policing (e.g. courts, formal institutions) negotiated contracts which encourage cooperative acts.

Up until 1950, economists conceded that no analytical model can predict the outcome of a negotiation. Bargaining was considered a bridge-too-far; a genuine realm of indeterminacy. Until, that is, Nash (1950, 1953) came along purporting to have pinpointed the *uniquely* rational agreement. To illustrate *Nash’s bargaining solution*, suppose Jill and Jack are negotiating over how to share an asset of value

V ; an asset that they can only enjoy if they manage to reach an agreement.

Two conflicting forces pull their bargaining behaviour in opposite directions: the *fear of impasse or conflict* (and, therefore, the loss of V for both) recommends a ‘softer’ negotiating stance, whereas the *fear of an inferior share* of V hardens their resolve.

Table 3.2

Rachel and Charles under a cloud of dominant belligerent strategies: R2 and C2 are best strategies for R and C regardless of what C and R will choose

Rachel’s (row) and Charles’ (column) strategy choices	C1: Non-violence	C2: Medium intensity conflict	C3: High intensity conflict
R1: Non-violence	4, 4	2, 5	0, 3
R2: Medium intensity conflict	5, 1	3, 3	2, 2
R3: High intensity conflict	3, 0	2, 3	1, 0

3.4.2 John Nash's solution to the bargaining problem

Nash began his analysis of bargaining by stripping it to its bare bones. He assumed that, after face-to-face negotiations that last for a pre-specified period, Jill and Jack retire to separate rooms where they cool off and, within another pre-specified period, write on a piece of paper their final claims over V : Jill claims x_L per cent of V and Jack x_K per cent of V . A 'referee' then collects their separate claims and sums them up. If $x_L + x_K \leq 100$, they both get what they claimed (the case of agreement). If $x_L + x_K > 100$ neither gets anything (the case of impasse). Do *uniquely rational* claims for Jill, say x_L^* per cent, and for Jack, say x_K^* per cent, exist? If they do, can Game Theory predict them? Nash (1950) proved that, under certain conditions, the answer is affirmative on both counts.

The proof begins with a model of Jill and Jack's behaviour borrowed in its entirety from neoclassical economics; namely, the model of an instrumentally rational agent whose behaviour succeeds in bringing about the outcome which corresponds to her maximum utility, given all her current constraints, some of which are due to what other people do.

In this context, Jill and Jack are assumed to derive utility from their shares of V and to care *only* about the size of their *own* share. Differences in their motivation are, naturally, catered for by assuming that they may value fractions of V differently or, equivalently, they may fear impasse differently.

Once Jill's and Jack's motivation has been defined, Nash (1950, 1953) shows that, *given some additional behavioural assumptions, there exists a uniquely rational agreement* (x_L^*, x_K^*) such that there will be neither impasse nor wastage (in short, $x_L^* + x_K^* = 100$) – for more detail on the assumptions necessary for the proof, see Hargreaves-Heap and Varoufakis (2004),

Chapter 4

. The gist of this agreement is simple: Nash predicts that Jill's share will be greater the less risk averse she is relative to Jack. Put differently, the more Jill fears impasse (relative to Jack), the less willing she is to risk bringing it on by demanding small increases in her portion of the 'pie' and, hence, the more prone she will be to settling for a (relatively) smaller share.

In its full technical version, Nash's proposed solution to the bargaining problem predicts that Jill and Jack will settle for an agreed distribution, such that the last fraction of Jill's share (i.e. of x_L^* per cent) yields a proportional increase in *her* utility *identical* to the proportional increase in Jack's utility caused by the last fraction of *his* share (i.e. of x_K^* per cent). It is fairly straightforward to show that this property of the proposed agreement is equivalent to suggesting that rational negotiators will settle on *a division that maximises the product of their utilities*.

The remarkable feature of Nash's solution is his claim that it constitutes *the uniquely rational outcome of bargaining*. It is one thing to suggest some way of

settling disputes and dividing pies; it is quite another to show that it is the *only* one that reason recommends. So, how did Nash prove that his proposed agreement is *the* rational one? A sketch of his proof follows:

Suppose that Jack offers Jill x_L per cent of the pie's value V but she rejects it demanding a higher share of, say, y_L per cent, threatening Jack that, unless he relents, she will abandon the negotiations with probability p . Jill's rejection is deemed *credible* if she prefers, on average, the prospect of getting y_L per cent of V with probability $1 -$

p rather than x_L per cent of the pie with certainty. Next, let us define some agreement A to be an *equilibrium of fear agreement* as follows: when Jill offers A to Jack, and he *credibly rejects* it in favour of some alternative division B , then Jill can *credibly reject* B (for *all* B) in favour of her original suggestion A . Nash first proves that bargainers will *only* settle for an *equilibrium of fear agreement* and then proves that there exists only one such agreement: his proposed solution (x_L^*, x_K^*) to the bargaining problem. *QED!*

Noting that the above proof applies for the general case of $N(>1)$ bargainers, it transpires that, in a few short pages of mathematical proof, Nash seems to have derived a definitive theory of mutually beneficial agreements between rational people with contradictory interests. Let us pause for a moment to contemplate the significance of this theoretical claim: Consider the foundations of any organisation, from a corporation, country club, a trades union, etc., to the melange of a society's legal and political institutions that determine the distribution of property and income, as well as the mechanisms for re-distribution which characterise contemporary states. Are they politically *legitimate*? Can they be ethically *justified*? The answer is affirmative if and only if there exists a rational agreement between their members to which they would converge as a result of a negotiation not dissimilar to that envisaged by J.-J. Rousseau in his *Social Contract* (1973).

Seen from the perspective of the present chapter's inquiry, what Nash is telling us is that, in all cases of antagonistic interests, there exists a uniquely rational agreement. The inference here is that, therefore, conflict can never be rational! Alas, Nash's logic is susceptible to a simple objection that brings it to its knees.

3.4.3 Nash's subterfuge

A pivotal aspect of Nash's logic, and of his proof, was the assumption *that a rejection is credible only to the extent that it is backed by a threat to cause conflict with probability $1 - p$* . Nash's theorem identifies an *equilibrium of fear agreement* with his bargaining solution and portrays the latter as a unique equilibrium of fear that one's (credible) rejection will be rejected (credibly). However, for this equilibrium to come about in practice, it must be the case that, while bargaining, Jill and Jack have *common knowledge* of the true value of p every time an offer is made or is turned down. But this is a tall order.

Indeed, for two people to labour under common knowledge of the outcome of $1 + 1$ is one thing; but to entertain commonly known subjective probabilities is

quite another. To have *common knowledge* that Jill will go to the movies tonight with probability 46.52% means not only that Jack predicts with 100 per cent certainty that Jill will go to the movies with probability precisely equal to 46.52% but, also, that Jill is 100 per cent sure that Jack is 100 per cent certain that Jill will go to the movies with probability 46.52% etc., etc. If such common knowledge sounds a little extreme, common knowledge of probability p in our analysis of bargaining above is utterly absurd.

For we know that Jill has good reason to under-play the true value of her p every time she rejects Jack's offer (since $1 - p$ is the threat of conflict with which she is trying to extract a concession from him). So, in a strategic environment in which players have strong incentives to shroud their p -choices in mystery, the assumption that these probabilities can be commonly known is impossible to digest. In conclusion, the extent to which one believes that Nash solved the bargaining problem coincides with one's readiness to accept that in the game of

Table 3.1

there exists only one rational course of action for each player because neither of them dares contemplate the idea that their beliefs are not transparent to one another. Then again, why should they not contemplate that idea? The only genuine answer is

because, if they do, economists will not be able to claim that they defeated the bargaining problem's ... indeterminacy!

3.4.4 Ariel Rubinstein's analysis of the bargaining process

Faced with this renewed encounter with indeterminacy, economists tried another tack. Having noted that Nash tried to do the impossible, in modelling the bargaining solution without modelling the bargaining process *per se*, they considered the possibility that a convincing theory of conflict might emerge if they modelled directly the bargaining process; that is, the sequence of offers, demands, counter-offers and counter-demands that is the drama of any negotiation.

To get gradually into the logic of forward looking into the future stages of the bargaining process, consider the simplest example: Jill is asked to suggest to Jack how to split \$100 between them. If Jack rejects her suggestion, the \$100 will shrink to a measly \$1 and it will be Jack's turn to offer Jill a portion of the remaining \$1. If she rejects it, no one wins anything. Backward induction, coupled with first-order commonly known rationality, leads us to the conclusion that Jill would make Jack an offer he could not refuse: 'You take \$1 and I keep \$99!' Now, let us consider a richer setting. Again Jill and Jack are given the opportunity to split \$100 with Jill making the first 'move.' Jack either accepts her offer or counter-proposes an alternative settlement. However, to add some urgency to the proceedings, let us imagine that, if Jack rejects Jill's initial offer, a timer starts ticking and, with every second that passes without agreement, 1c is taken off the \$100 prize. That is, if they take M minutes to reach agreement, the \$100 will, by then, have shrunk to $\$(100 - 0.6M)$.

How should one play this game? Jill must now balance the urge to make Jack an offer that he will not refuse (so as to avoid 'shrinkage' of the prize) against the

worry that she might end up offering him too much. Recall that, in all bargaining games, *any* outcome is rationalisable (moreover, any outcome is a Nash equilibrium). If, for example, Jill expects Jack to accept 40 per cent and thus issues a demand for 60 per cent, while Jack anticipates this, then a 60–40 split is an equilibrium outcome (as it confirms each bargainer's expectations). And since any outcome is rationalisable, the theory offers no guidance to players. To the rescue comes the method of working out backwards what will happen at each stage of the negotiation (beginning at the last stage, moving to the penultimate one and so forth) while, at the same time, assuming that the bargainers' strategies will be in a Nash equilibrium at each stage. Let us refer to this analytical method as *Nash backward induction*.

Consider the following strategy that Jack may employ in his negotiations with Jill: 'I shall refuse any offer that awards me less than 80 per cent.' This may be rationalisable (and a Nash equilibrium) when we look at the final outcome independently of the bargaining process, but it may not be if we examine the various alternative strategies against the background of the actual bargaining process. Why? Because such a strategy may be based on an *incredible threat*. This is why:

Suppose Jill offers Jack only 79.9%. Were Jack to stick to his 'always demand 80 per cent' strategy, he would have to reject the offer. However, this rejection would cost him as the prize shrinks continually until an agreement is reached. Even if his defiant strategy were to bear fruit soon after the rejection of Jill's 79.9% offer (i.e. if Jill were to succumb and accept Jack's 80 per cent demand M minutes after her 79.9% offer was turned down), Jack will only get 80 per cent of a smaller prize. How much smaller the prize will be depends, of course, on M ; i.e. on how long it will take Jill to accept Jack's demands. If it takes more than 12.5 seconds, Jack will be worse off than he would have been had he accepted her offer of 79.9%!

9

Thus, if it is commonly known that it takes well over ten seconds for bargainers to respond to an offer, Jack has no incentive to stick to the strategy 'always demand 80

per cent.’ And so, if during negotiations Jack threatens to reject *any* offer less than 80 per cent, Jill should take this threat with a pinch of salt; and a very large one if it takes more than about 10 seconds to make a response to any offer.

The above is an important thought. By means of Nash backward induction, we can discard a very large number of possible negotiating strategies on the basis that they will not work if the agents’ rationality is commonly known. Ariel Rubinstein (1982) used this logic to prove a remarkable theorem: There exists only *one* equilibrium that does *not* involve use of incredible threats. The brilliance of this thought matches that of John Nash’s original idea for solving the bargaining problem and, what is even more extraordinary, yields a solution analytically equivalent to that of Nash as the time delay between offers and demands tends to zero (the latter was shown by Binmore, Rubinstein and Wolinsky, 1986).

3.4.5 A proof of Rubinstein’s theorem

The precise bargaining process examined by Rubinstein is very similar to the preceding example. There is a prize to be distributed and Jill kicks the process off by making a proposal. Jack either accepts or rejects it. If he rejects, it is his turn to make an offer. If, in turn, Jill rejects that offer, the onus is on her to offer again, and so on. Every time an offer is rejected, the prize shrinks by a certain proportion which is called the *discount rate*. Analytically it is very simple to have different discount rates for each bargainer and this allows one to introduce differences between the bargainers, differences that are equivalent to the differences in the rates of change of utility functions (or *risk aversion*) discussed earlier in the context of the Nash solution. Rubinstein’s theorem asserts that rational agents will behave as follows: *Jill will make Jack an offer that he cannot refuse* (or, more precisely, *does not want to refuse irrespectively of how much he likes it*).

Thus, there will be no delay and the prize will be distributed before the passage of time reduces its value. Moreover, the settlement will reflect two things:

- (a) Jill’s first-mover advantage, and
- (b) Jill’s relative eagerness to settle (i.e. their relative discount rates).

By (a) we imply that Jill (who makes the first, and allegedly, final offer) will retain (other things being equal) a greater portion than Jack courtesy of the advantage bestowed upon her by the mere fact that she offers first (something like the advantage of the white player in chess). [Note, however, that if offers can be exchanged *very* quickly, the first-mover advantage disappears (in the limit).

¹⁰

] By (b) it is meant that eagerness to settle is rewarded with a smaller share. If Jack is more eager to settle than Jill, then he must value a small gain now more than Jill does, as compared with a greater gain later.

This result is perfectly compatible with Nash’s solution which, as we have shown, penalises *risk aversion*. To the extent that *risk aversion* and an *eagerness to settle* are similar, the two solutions (Nash and Rubinstein) are analytically interchangeable. This is Binmore *et al.*’s (1986) contribution: they prove that, when agents exchange offers at the speed of light, and their discount rates reflect their risk aversion, Rubinstein’s solution is identical to that of Nash.

Let us now look at the proof once we have defined the player’s discount rates and a statement of the theorem. Every time an offer is rejected, Jill’s valuation of the prize loses a proportion given by $1 - \alpha$ (where α lies between 0 and 1). It is as if, in her eyes, portion $1 - \alpha$ of the pie has been lost. Similarly, with every rejection that occurs, Jack’s valuation of the prize diminishes by $1 - \beta$. For example, if $\alpha = \beta = 0.8$, then, when an offer is rejected, only 80 per cent of the prize is preserved in the next round. Thus, if Jill and Jack come to an agreement at $t = 3$, the prize they will be splitting will have shrunk twice; the extent of the ‘shrinking’ depends on α and β .

¹¹

These parameters (α and β) are known as the bargainers’ *discount rates*. They

are closely related to the player's *risk aversion*. To see this consider the position of the person deciding at some stage whether to accept the other's offer.

The player has a choice between accepting a share of the cake now or rejecting the offer and bargaining over the pie in the next time period. The outcome of the bargain in the next time period is uncertain whereas acceptance of the offer now yields a known quantity. The extent of the perceived shrinkage of the pie will then reflect the person's perception of the risk associated with this uncertainty.

Discount rates α and β are also sometimes known as the bargainers' *time preferences*; referring to their capacity to capture the players' valuation of a larger payoff tomorrow compared to a smaller one today. By comparing discount rates we gauge the bargainers' relative urgency to settle. For example, if $\alpha > \beta$, Jill is clearly less eager to settle than Jack (as the prize shrinks, with every failed offer, (relatively) faster for him than it does for her). In any case, it is evident that the ratio α/β is a good proxy for Jill's relative fear of disagreement (as compared to Jack's). For if $\alpha/\beta > 1$, Jill loses less from each rejection (and, thus, from each delay in reaching agreement) than Jack. Other things being equal, we might therefore expect Jill to be less acquiescent to Jack the higher the value of α/β . In this dynamic sense (that is, when time comes into the bargain), ratio α/β is the equivalent to the relative risk aversion that determined the outcome in Nash's 1950 solution.

We now set to prove Rubinstein's theorem which states that, at the very outset (i.e. at $t = 1$), Jill will make Jack an offer that he will accept immediately. That is, there will be no conflict whatsoever, as long as they are rational! And what will that offer be? Rubinstein's answer is: Jill will proposed a split of the pie along the division

$\left(\frac{1-\beta}{1-\alpha\beta}, \frac{\beta(1-\alpha)}{1-\alpha\beta} \right)$. Let us see how this remarkable result can be proven.

At this stage, I must warn the reader that the proof relies heavily on a *hidden assumption*. While postulating an indefinite bargaining horizon, Rubinstein presumes (without stating this explicitly in his paper) that there shall come a stage, call it round k , at which Jill's and Jack's beliefs (regarding the maximum share they can each expect to get) will have converged. Furthermore, Rubinstein's hidden assumption has it that Jill and Jack have common knowledge (even at time $t = 1$) of that distant round k . This is the hook on which Rubinstein secures the logic of *Nash backward induction*. Thus the latter unravels, beginning at $t = k$, then moving to $t = k - 1$ and finally to $t = 1$ where Jill's unique equilibrium offer to Jack is computed. And since it is a unique equilibrium offer, Jack simply accepts it. In summary, our proof will involve four steps:

- (State the *hidden assumption* that gives Nash backward induction its foothold.
- A)
- (Consider the minimum offers Jill and Jack will issue at $t = k, t = k - 1, t = k - 2, \dots, t = 1$. [I also prove that they have no incentive to offer less than these
- B) minimum offers (namely, that their minimum offers equal their maximum offers).]
- (Once the unique offer at $t = 1$ is derived, we utilise (again) the *hidden*
- C) *assumption* to argue that Jill must propose a division at $t = 1$ identical to that
- (which they would agree on (at much greater cost) at $t = k$. Once this assumption
- (is made, Jill's offer to Jack at $t = 1$ will be computed.
- (Prove that the actual value of k does *not* matter, as far as the bargaining
- D) solution is concerned. All that matters is that, in accordance to the *hidden*
- (*assumption*, we presume that at $t = 1$ bargainers entertain common knowledge of
- (k (whatever its value might be).

Step A: The hidden assumption

There are two parts to this assumption: (a) There exists some round $t = k (>2)$ in which

bargainers' beliefs will have become consistently aligned on distribution $(V, 1 - V)$, and (b) Jill and Jack have *common knowledge* of k at $t = 1$. The *hidden assumption* is, of course, a reincarnation of the earlier assumption that no 'false' beliefs are allowed (i.e. only consistently aligned beliefs are permitted). Part (b) is familiar territory: If there exists an unknown parameter (k in our case) and players are equally rational and well informed, their estimates of k must be common (and it must be commonly known that they are common). Part (a) is another application of the banishment of erroneous beliefs: Players assume that their beliefs about the outcome will be, at some stage (k), consistently aligned.

Thus, Rubinstein has the anchor for the logic of Nash backward induction that he needed: the k th round of the bargaining process. From there one simply moves backwards through rounds $k - 1$, $k - 2$, ... and, finally, to the first round.

Step B: Computing bargainers' offers backwards

Consider the value $k = 3$. Why $k = 3$? Because it is a small number of stages which will help us keep the proof simple. Do we not lose generality by assuming such a small number of stages? No, because as we shall see in Step D, the actual choice of k makes no difference to the proof. And since it makes no difference to the proof, or to the bargaining solution, whether k equals 3 or 3,000 we might as well keep things simple.

So, suppose that Jill and Jack commonly know at $t = 1$ that by stage $t = k (= 3)$ they will discern the *same* solution, say $(V, 1 - V)$, to the negotiations over the distribution of the pie. This does not, of course, mean that at $t = 1$ they know the value of V which will surface at $t = k = 3$. All it means is that at $t = 1$ they have common knowledge of the 'fact' that, come $t = k = 3$, there will be *some* portion of the pie, say V , which (in the eyes of both Jack and Jill) Jill will not be able (or willing) to improve upon (by means of more bargaining).

Table 3.3

acknowledges this in its first row, which depicts the bargainers' offers and demands in round $t = k = 3$. Once more, let us remind the reader that neither we (as theorists), nor our bargainers have any way (at this stage of the theorem's proof) of knowing the value of V . All that is known at $t = 1$ is that, come $t = 3$, both will have in their minds *some* value V which will be a commonly held estimate of Jill's final share of the pie. The task is to compute this value.

Let us now investigate Jack's situation at $t = 2$. It is his turn to accept or reject the offer Jill made him at $t = 1$. Should he reject Jill's $t = 1$ offer (and come up with a counter-offer)? Or should he accept it? If he rejects it, what counter-offer

should he make? He knows (from the *hidden assumption*, with $k = 3$) that, if the negotiations proceed to $t = k = 3$, Jill can expect to get V . So, Jack knows that if he were to offer her, at $t = 2$, portion αV of the pie, she has no reason to turn it down: indeed, she cannot reasonably expect (given the *hidden assumption*) to do better. The reason, of course, is that Jill's $t = 2$ valuation of V at $t = 3$ equals α (her discount rate) times V . Put differently, Jill must (by definition) be indifferent between portion αV at $t = 2$ and V at $t = 3$. Any offer by Jack (to Jill) at $t = 2$ below αV will spark off a rejection.

Table 3.3

The backward induction of optimal offers based on the hidden assumption

Round	Proposer	Proposed share for Jill	Proposed share for Jack
$t = k = 3$	Jill	V	$1 - V$
$t = 2$	Jack	αV	$1 - \alpha V$
$t = 1$	Jill	$1 - \beta(1 - \alpha V)$	$\beta(1 - \alpha V)$

The *hidden assumption* asserts that there is some commonly known round ($t = k$) during which Jill and Jack will believe that the pie will be divided in portions ($V, 1 - V$). Round k is commonly known at all stages. The path of the Nash-backward-induction is depicted by the direction of the arrows. Supposing that $t = k = 3$, at $t = 3$ (if the negotiations last that long) Jill will receive portion V . Thus, at $t = 2$ Jack must offer her αV to avert conflict, keeping $1 - \alpha V$ for himself. Thus, at $t = 1$ must offer him at least $\beta(1 - \alpha V)$ so as to avoid a rejection that will delay agreement. If she does this, she keeps $1 - \beta(1 - \alpha V)$.

Thus we computed Jack's minimum offer to Jill at $t = 2$ if he wants to avert a rejection by Jill: It is an offer of portion αV . The question now becomes: Will Jack want to avert a rejection at $t = 2$? If he offers Jill anything below αV , Jill will reject it and bargaining will continue until $t = 3$ where Jill will receive portion V and Jack $1 - V$. Jack knows this at $t = 2$. Thus he knows that, if he offers Jill less than αV , the most he can get at $t = 3$ is $1 - V$. What is $1 - V$ at $t = 3$ worth to Jack at $t = 2$? Since his discount rate is β , $1 - V$ at $t = 3$ is worth to Jack $\beta(1 - V)$ at $t = 2$.

In short, Jack has a stark choice at $t = 2$: Offer Jill αV immediately (at $t = 2$); an offer that we know she will accept, therefore leaving him at $t = 2$ with $1 - \alpha V$ of the pie. Or, offer her less than αV ; a move that will lead him to a payoff whose value to him, at $t = 2$, equals $\beta(1 - V)$. Clearly, he will induce rejection at $t = 2$, by offering Jill less than αV of the pie, only if $\beta(1 - V) > 1 - \alpha V$. However, as all these parameters (α, β and V) lie between 0 and 1, this inequality is *never* satisfied. Which means that, at $t = 2$, Jack will *never* offer Jill less than αV . And since (as we have already shown) an offer of αV is the minimum she will accept, Jack has no reason to offer her more than that. Thus we have proved that Jack's maximum offer at $t = 2$ will be the minimum that Jill will accept and we have a

uniquely rational offer: At $t = 2$ Jack offers Jill portion αV and she accepts it. We make a note of this result in the second row of

Table 3.3

With the analysis of round $t = 2$ complete, we now turn our attention to what happens during $t = 1$ at which point it is Jill's turn to make an offer. She knows (see above) that if her offer is turned down, bargaining will proceed to round $t = 2$ where she will be offered portion αV ; an offer that she will accept. By the same token, she knows that Jack knows that he can expect, at $t = 2$, a sure portion of $1 - \alpha V$. What is his valuation at $t = 1$ of portion $1 - \alpha V$ at $t = 2$? Given his discount rate of β , at the outset ($t = 1$) Jack's valuation of his share of the second round ($1 - \alpha V$) equals $\beta(1 - \alpha V)$. Thus we (and Jill along with us) know that if Jack is offered portion $\beta(1 - \alpha V)$ at $t = 1$, he will accept it. Any offer less than that will cause disagreement and a counter-offer by Jack at $t = 2$. The question then, predictably, becomes: Does Jill want to settle with Jack at $t = 1$?

If she offers Jack less than portion $\beta(1 - \alpha V)$, he will reject her offer and he will return to the bargaining table with the suggestion that Jill keeps portion αV ; an offer that, as we have already proven, Jill will accept. So, it all comes down to whether Jill

prefers $1 - \beta(1 - \alpha V)$ immediately (i.e. at $t = 1$) or the prospect of a certain portion equal to αV in the next round ($t = 2$)? Since the later is valued by Jill at $t = 1$ at $\alpha^2 V$, she will opt for immediate settlement (at $t = 1$) if $1 - \beta(1 - \alpha V) > \alpha^2 V$.

It is easy to show that this inequality holds *always*! This means that Jill will *always* prefer to offer Jack portion $\beta(1 - \alpha V)$ at $t = 1$; an offer that he has *no* incentive to reject and which Jill prefers to offer over any alternative offer that will cause Jack to turn it down. We have, consequently, reached the conclusion that at $t = 1$ Jill will offer Jack division $[1 - \beta(1 - \alpha V), \beta(1 - \alpha V)]$. And Jack will accept it.

Step C: Consistent preferences over time

Before we investigate further, consider any case of conflict between two persons, countries, firms etc. If they both knew at the outset *how* the 'war' between them would be settled, would it not be rational to agree at the very beginning to settle it in that manner while skipping the costly fighting? In our case this would mean that Jill tells Jack: 'We know (recall the *hidden assumption*) that if we wait till $t = k$, I shall receive V portion of the pie. Why wait until then? Let me have portion V now and, in this manner, no part of the pie will be lost (through delay in reaching agreement) for either of us.'

Rubinstein assumes that an instrumentally rational Jack has no reason to disagree; and we call this the assumption of *consistent preferences over time*. The only problem is that they do not know the precise value of V . However, there is a way of discovering it now. We have concluded above that both will entertain the same expectation of what Jill will get if they ever reach $t = 3$: Jill will get V . At the same time, we have concluded that, at $t = 1$, Jill will demand $1 - \beta(1 - \alpha V)$ for herself and Jack will let her have it. So, why not say that she will demand now

$[1 - \beta(1 - \alpha V)$ at $t = 1]$ the same share (V) that she will get if she were to hold out until $t = k = 3$? In other words, the assumption of consistent preferences over time,

implies that $1 - \beta(1 - \alpha V) = V$. Solving this simple equation for V , we find $V = \frac{1-\beta}{1-\alpha\beta}$.

This completes the proof of Rubinstein's theorem according to which Jill and Jack will settle at $t = 1$ on portions $V = \frac{1-\beta}{1-\alpha\beta}$ and $\frac{\beta(1-\alpha)}{1-\alpha\beta}$ respectively. Why at $t = 1$? Because, as rational people, they recognise that the equilibrium division will be the same whether they settle immediately or much later and, therefore, conclude that there is nothing to gain (and much to lose) from delaying the agreement.

Step D: k does not matter

The proof above relies on the assumption that $k = 3$; namely, that it is common knowledge to Jill and Jack (at $t = 1$) that in the space of merely three periods their beliefs on the outcome will have become consistently aligned. We shall now show that this was assumed only for convenience as the proof of Rubinstein's theorem holds for any finite value of k . To get a flavour of why this might be so, let us consider the case $k = 5$.

Table 3.3

is now replaced by

Table 3.4

in which there are an extra two rows and backward induction begins at $t = k = 5$. Otherwise, the three first rows of

Table 3.4

are identical to

Table 3.3

Applying the logic of consistent preferences over time (as we did in the previous stage of the proof),

12

Jill is assumed to make a demand at $t = 1$ equal to the share of the pie she can

expect to get were she to hold out until $t = k = 5$. In other

words, $V = 1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$. Solving for V we find Jill's optimal opening demand of $V = \frac{1-\beta}{1-\alpha\beta}$; precisely the same offer as we had when $k = 3$. More generally, for any value of k , Step D yields the following equation:

$$V = 1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha(1 - \beta(1 - \dots\alpha V))\dots)]\}$$

Table 3.4

The case of $k = 5$

Round	Proposer	Proposed share for Jill	Proposed share for Jack
$t = k = 5$	Jill	V	$1 - V$
$t = 4$	Jack	αV	$1 - \alpha V$
$t = 3$	Jill	$1 - \beta(1 - \alpha V)$	$\beta(1 - \alpha V)$
$t = 2$	Jack	$\alpha[1 - \beta(1 - \alpha V)]$	$1 - \alpha[1 - \beta(1 - \alpha V)]$
$t = 1$	Jill	$1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$	$\beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$

We begin at the last stage ($t = 5$) where Jill gets V ; then we move to $t = 4$ where Jack offers her αV , keeping $1 - \alpha V$ for himself; then to $t = 3$ where Jill must offer him $\beta(1 - \alpha V)$, claiming $1 - \beta(1 - \alpha V)$ for herself; then to $t = 2$ where Jack offers Jill $\alpha[1 - \beta(1 - \alpha V)]$ to induce a settlement, claiming $1 - \alpha[1 - \beta(1 - \alpha V)]$ for himself; and, finally, to $t = 1$ where Jill offers Jack $\beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$, demanding $1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$ for herself. Setting $V = 1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha V)]\}$ and solving for V yields the same solution as in

Table 3.3

$$V = 1 - \beta\{1 - \alpha[1 - \beta(1 - \alpha(1 - \beta(1 - \dots\alpha V))\dots)]\}$$

Solving for V yields, as before, $V = \frac{1-\beta}{1-\alpha\beta}$ independently of how many extra 'stages' the dots (...) entail. In conclusion, as long as the *hidden assumption* holds, the actual value of k does not make a difference to the Rubinstein solution.

Recapping, the above proof shows that, in the context of the assumptions made, there is only one rational bargaining strategy that does not involve incredible threats: that is, there is one equilibrium. Of course, there are logical difficulties not only with the extra assumptions made (primarily the *hidden assumption*) but also with the use of Nash backward induction in the construction of the equilibrium outcome. The point here is that the Rubinstein solution is internally consistent, provided one assumes that out-of-equilibrium behaviour is explained by random trembles. If any deviation from the behaviour proposed by Rubinstein (e.g. rejection of Jill's demand V by Jack at $t = 1$) is interpreted by Jill as a random error, then Jill will take no notice of this rejection. And if it is common knowledge that Jill will take no notice of such a deviation from the equilibrium strategy at $t = 1$, then Jack cannot entertain rational hopes that by rejecting offer $1 - V$ at $t = 1$ he will bring about a better deal (e.g. $1 - W > 1 - V$) for himself. But why should one assume this? Why is it uniquely rational for Jill to see nothing in Jack's rejection at $t = 1$ which can inform her about his future behaviour? And why does Jack have to accept that Jill will *necessarily* treat his rejection as the result of a random tremble, rather than as a signal of a defiant, purposeful, stance?

Of course, it is entirely possible that Jill will not 'read' anything meaningful in Jack's

resistance to V at $t = 1$. It is equally possible that Jack will have anticipated this, in which case he will not reject $1 - V$. But, equally, it seems difficult to rule out, through an appeal to reason alone, the possibility that Jill will take notice of Jack's rejection of $1 - V$ at $t = 1$ and to see in it evidence of a 'patterned' deviation from Rubinstein's solution. If this happens, she may rationally choose to concede more to Jack. And if Jack has anticipated this, he will have rationally rejected $1 - V$ at $t = 1$. In conclusion, an equilibrium solution (like that by Rubinstein) may or may not hold ... rationally. Thus, the bargaining problem remains indeterminate. We simply have no clue, even after all these laborious computations, of what rational bargainers will do. Of whether they will settle without conflict or whether they will inflict mutual damage upon one another.

3.4.6 Objections to Rubinstein

Rubinstein's solution to the bargaining problem depends on the equilibrium method allied to three important, albeit potentially controversial, assumptions:

- (a) A further application of the assumption of no erroneous beliefs (i.e. of perfectly and consistently aligned beliefs) according to which both players know that, at a commonly known date k , they would settle for V and $1 - V$ respectively,
- (b) The assumption of consistent preferences over time,
- (c) The assumption that the rate of discount remains the same for each player over time.

We have already discussed the dire objections to the idea of consistently aligned beliefs. Perhaps the only thing we need add here is a comment on the innovative manner in which they were utilised by Rubinstein. By assuming consistently aligned beliefs on the number of rounds (k) it would take our bargainers to form consistent estimates on how the pie will be distributed between them (i.e. of division V and $1 - V$), Rubinstein cunningly introduces a 'final' stage of the bargaining game (stage k) which gives *Nash backward induction* the foothold it needs in some future date before it starts unfolding backwards (from $t = k$ to $t = k - 1$, to $t = k - 2$, ... to $t = 1$).

To make the same point slightly differently, the innovation in question is that Rubinstein uses consistently aligned beliefs in order to impose a finite end-state to an otherwise infinite-horizon dynamic game. Secondly, given the fixity of k , he puts it to work to constrain the bargainers' beliefs from straying off the equilibrium path. Those sceptical of constantly consistent beliefs should take note of the twin use to which it must be put before Rubinstein's solution to the bargaining problem is entertained.

Turning now to the other two assumptions underpinning Rubinstein's solution [(b) and (c) above], while they may sound like straightforward consistency requirements, they abstract from the common human experience of preferences that can be endogenous to bargaining. Thus, they ignore the possibility that people pay decreasing attention to material (i.e. money) payoffs and, instead, as the bargaining process unfolds (especially when their opponents prove more recalcitrant than expected), place more emphasis, for example, on 'beating' them. This psychological interplay is ruled out by Rubinstein (and, in all fairness, by all game theory). (I return to this issue in

Chapter 8

.)

To make our critique more concretely, let us use an example which helps bring out the problem of interpreting out-of-equilibrium bargaining behaviour. Suppose that Jill gets the bargaining process going and that $V = 0.6$. Jack's best strategy (according to Rubinstein's theory) is to accept 40 per cent of the pie instantly. What will happen if he rejects this and counter-claims, say, 60 per cent at $t = 2$? For this bargaining strategy to make sense, two conditions must hold: (a) there must exist a portion $W(> 0.4)$ of the pie which at $t = 2$ is worth more to Jack than 40 per cent of the pie did at $t = 1$; and (b) Jack must have a rational reason for believing that it is possible to get at least W at $t = 2$ if he rejects offer V at $t = 1$ and counter-proposes that he keeps 60 per cent.

Condition (a) is easy to satisfy provided the rate at which the pie is shrinking (in Jack's eyes) is not too high. Condition (b) is far trickier. Specifically, it

requires that the experience of an unexpected rejection by Jack may be sufficient for Jill to panic and make a concession not predicted by Rubinstein's model. This development would resemble a tactical retreat by an army which realises that, in spite of its superiority, the enemy may be, after all, determined to die rather than (rationally) withdraw; so it is not completely implausible. If Jack's rejection of offer $1 - V$ at $t = 1$ inspires this type of fear in Jill, then she may indeed make a concession beneficial to Jack; and if Jack manages to bring this about by straying purposefully from Rubinstein's equilibrium path, then it is not irrational to stray in this manner.

13

So, why are economists so determined to rule out that rational bargainers may stray from the equilibrium path? The only plausible answer is, of course, that unless they stick to this path, their analysis is lost in a forest of indeterminacy. Well, that is the economists' tragedy, as opposed to any reason to think that bargainers will stick to Rubinstein-like equilibrium bargaining behaviour!

3.4.7 On the impossibility of a neoclassical explanation of rational conflict: A remarkable theorem by F. Gul and H. Sonnenschein

Setting aside, for the moment, the argument that conflict is probably due to the fact that bargainers have no reason to stick to the path of equilibrium strategies, let us ask the following question: What explanation of conflict do neoclassical economists give, in view of their determination to assume (against Reason) that players will stick to equilibrium behaviour?

The answer any self-respecting neoclassical economist will give is: conflict is the result of asymmetrical information. That is, even when bargaining strategies are in equilibrium, neoclassical bargaining theory *can* predict a delayed agreement (which, when we assume that time costs money, is the equivalent of conflict) as a repercussion of an asymmetrical initial distribution of information among the antagonists.

Indeed, arms negotiators (especially during the Cold War) would confirm that the greatest hindrance to a convergence of views is the mutual suspicion which feeds on ignorance of the other side's motivation. In an environment of endemic uncertainty, no one can devise a unique strategy for rational bargainers and conflict may thus ensue. The economic literature is saturated with models allegedly dealing with the effects of uncertainty. We may not know exactly what to expect, they postulate, but we are assumed to know all eventualities in advance and, moreover, we can readily assign probabilistic expectations to each one of them. A tall order indeed! The fact that the future may bestow a phenomenon that was not considered at all (not the same as being considered and assigned a zero probability) is ignored. As other kinds of uncertainty cannot be handled by the tools of neoclassical economists, they treat them as ... inadmissible.

Faced with uncertainty, game theorists employ the standard-issue tools and proceed surgically to remove informational disorders. Take the Nash bargaining model, for example (see

[Section 3.4.2](#)

), and suppose that Jill does not know Jack's utility function. In the simple variant of the bargaining game, where both sides

make a secret bid for the portion of the pie that should go to them, one would expect all sorts of possible outcomes; that is, indeterminacy. Indeed, it would be a miracle if the two bids or demands summed up precisely to the slice of the pie. And yet game theory has found a way to show that the Nash bargaining solution does not lose its power once we introduce uncertainty.

The major contribution in this area is that of Harsanyi and Selten (1972). They ask

us to suppose that Jack and Jill may be one of two types of person: 'hard' or 'soft', each with different utility functions: u_L^h and u_L^s for Jill and u_K^h and u_K^s for Jack. Provided that these functions are commonly known, all that is needed for the solution of the bargaining game is that Jill must harbour some probabilistic expectation of the type of person Jack is, and vice versa. The bargaining game is no longer played by the original two players, Jill and Jack, but by their possible selves, each being given a weight proportional to how likely her or his opponent thinks it is that 'it', the self, is the true one.

Imagine there are two rounds. In the first, a random draw decides whether Jill and Jack will be 'hard' or 'soft' with probabilities m and n respectively. Both players know the probability that their opponent is of the 'strong' or 'soft' but do not observe which type was selected in the first round. Thus, each knows with certainty who he or she is ('hard' or 'soft') but only has a probabilistic expectation of what their opponent is. In the second round, each side makes the usual demand or bid regarding the portion of the pie they want.

Harsanyi and Selten (1972) then show that the resolution of this two-person game, under uncertainty, is equivalent to a bargaining game between Jill's and Jack's four possible selves, where the 'power' of each of these selves is proportional to how probable it is that they are the true self of Jill or Jack. The gravity of the preferences of each potential type of player is increasing with the probability that they were selected, in the first round, as the player's true self or type. At the end of the second round, the Nash bargaining solution x^* (where x is the portion of the pie that goes to Jill and $1 - x$ is Jack's portion) maximises the product of the roles utility functions, suitably weighted by how probable their selection was in the first round. Thus, the uniquely rational agreement (according to Nash, as extended by John Harsanyi and Reinhard Selten) is given as:

$$x^* = \operatorname{argmax}\{(u_L^h(x))^m (u_L^s(x))^{1-m} (u_K^h(1-x))^m (u_K^s(1-x))^{1-m}\} \quad (3.1)$$

Now, recall that in the standard, symmetric bargaining problem, Nash's solution to the bargaining problem is $x^* = \operatorname{argmax}\{u_L(x)u_K(1-x)\}$. Then, suppose that Jill and Jack have different bargaining powers. The Nash solution would need to be re-written as:

$$x^* = \operatorname{argmax}\{u_L(x)^\lambda u_K(1-x)^\mu\} \quad (3.2)$$

It is now evident that the weight (m or n) attached in the shared information game to one's utility function reflects one's power. It is, in fact, the distribution of information in the above bargaining game that determines (at least partially) the distribution of power. In (3.2) the larger the ratio μ/λ the greater Jill's payoff. So, in (3.1) the larger the ratio of m/n the greater Jill's payoff *even if she is 'soft'*. Having a reputation for toughness is almost as important as being tough. With this commonsense result behind us, let us now see what effect imperfect information has on conflict.

Harsanyi and Selten (1972) distinguish between two kinds of equilibrium solutions. The first is one in which all players act in an identical manner. As this does not allow an observer to deduce their true 'type', simply by observing their behaviour, they are called *non-revealing equilibria*. Solutions in which it is optimal for different types to do different things are thus *revealing*. Clearly, the latter case, of *revealing equilibria*, is much more interesting. The question then becomes: When is it rational for Jill to behave differently when she is 'soft'? Is there any sense in revealing that she is not a 'hard' bargainer? Or does she always have an incentive to bluff? To pretend that she is a 'tough cookie' when, in reality, she is a marshmallow?

To explore these questions, let us ask a fresh one: How much power would Jill need to have to extract from Jack the same share in this game as she would have had her true identity been observable by him? This value can be easily computed by comparing weight m in (3.2) to μ in (3.1). Given this power estimate, Jill would prefer information to be shared with Jack if the equivalent bargaining power that she has under asymmetric information (m) is less than that under perfect information. In a repeated

bargaining interaction, however, there may emerge a credibility problem. How can the 'hard' Jill inform Jack of her true nature when Jack thinks of her, mistakenly, as 'soft' with probability $1 - m$? Might it not be the case that Jill must destroy something of value to herself, prior to issuing her demands, in order to signal what 'she is made of'? Is this not the way to explain rational conflict analytically?

Not necessarily. Senseless destruction is not in itself a persuasive signal that will make Jack want to roll over. Before he sets his probabilistic assessment that Jill is truly 'hard' equal to one, he must observe behaviour that a 'soft' Jill would consider more expensive than being revealed to be 'soft.' On the other hand, a 'hard' Jill may not always play tough. If the difference between the 'hard' and 'tough' roles is small relative to the difference in costs that must be incurred in order to convince Jack, then the price of convincing him that she is 'hard' may not be worth paying – even if she truly is 'hard.'

We are beginning to discern a potentially clever analysis of rational conflict which portrays conflict as a possible signal that helps agents identify their true nature. After the destructive action has been taken by Jill, Jack updates his estimation that she is 'hard.' He does this after attempting to replicate the thoughts of a weak Jill. For if the same tough choice could be made by a 'soft' Jill, quite rationally and profitably, observation of conflictual tendencies by Jill would not reveal to him anything he did not already know.

Suppose that the 'soft' Jill can look forward to portion y of the pie without playing tough, whereas she will receive x if she convinces Jack that she is 'hard' ($x > y$). If c_s is the cost of conflict for the 'soft' player, the relevant equilibrium

solution will be of the revealing kind when $x - c_s < y$. This inequality guarantees that a 'soft' Jill has nothing to gain from opting for conflict. Which means that, if she is observed choosing conflict, she *must* be 'hard' (or irrational) as long as $x - c_h > y$, where c_h is the 'hard' Jill's conflict cost. Hence, the first glimpse of rational conflict theory seems to be a stone's throw away and promises an explanation of why rational people fight along the lines of: 'Because conflict may be a credible signalling mechanism that restores informational equity.'

Summarising, in the case of such revealing equilibria, conflict seems to have been rationalised in a bargaining game in which bargainers cannot, in truth, exchange demands and offers at will. There is an initial round in which some randomisation determines the type of each player, another one in which they choose to destroy or not to destroy a valued resource (e.g. part of the pie) and, finally, a round in which they submit sealed bids for the (remaining) pie. If the sum of these bids equals to or is less than the pie, they get what they demanded. Otherwise neither gets anything.

Can this simple framework be extended to a fully fledged model of rational bargaining where conflict is one possible outcome of asymmetrically distributed information? Was John Hicks (1932) wrong when he argued that the outcome of bargaining is necessarily indeterminate and that this is the reason why conflict may indeed result when rational people tussle and bargain with one another? A paper in 1985 by Ariel Rubinstein seemingly answered these questions in the affirmative: Rubinstein (1985) is a revamped version of the original 1982 model which allows for uncertainty regarding your opponent's discount factor; i.e. degree of patience. In this context, where there is no bound to the number of potential bargaining rounds, Rubinstein shows that a small amount of delay in reaching an agreement (as opposed to the instantaneous agreement that his 1982 model demands) can be explained as the price bargainers may have to pay in order to signal to each other their true degree of patience or urgency to settle. Thus, the Harsanyi and Selten (1972) model was extended to sequential bargaining rounds that can go on and on, as long as Rubinstein's hidden assumption lurks in the model's shadows; that is, as long as

common knowledge of the future period in which there is probability one of a shared belief on the final agreement.

Three years later, Gul and Sonnenschein (1988) came to explode this claim. For they showed that conflict disappears from *any* bargaining model that respects the neoclassical (and game-theoretical) edicts, provided bargainers are not constrained by the theorist with regard to when and how they will alter their offers and demands. Recall how in Harsanyi and Selten (1972) Jill and Jack were severely constrained to issue only one bid or demand. In more complex variants of this type of model, such as that of Rubinstein (1985), theorists assume that we can have a large – and even an unspecified – number of bargaining rounds (during which players can update their offers and demands), separated, however, by a fixed time interval during which no offers and demands can be altered. If this time interval is allowed to tend to zero (that is, if players are allowed to send text messages to one another at any point in time, renewing their offers or demands at will and in continuous-real time), suddenly there are no more discrete rounds. Gul

and Sonnenschein's (1988) remarkable theorem shows that, even when information on the types of bargainers is asymmetrically distributed, the delay in reaching agreement tends to zero if time flows continuously (as opposed to discretely). The possibility of rational conflict thus disappears!

The significance of this result for our inquiry cannot be overstated. For here is the dilemma: If we insist on the equilibrium logic of Nash and the neoclassicists, we cannot have a theory of why conflict is observed when rational agents bargain with one another. As per the Gul and Sonnenschein theorem, conflict periods vanish as bargainers are allowed full freedom to exchange offers and demands at will. If, on the other hand, we permit bargainers to issue demands and offers that lie off the equilibrium path, then rational conflict becomes possible again but, suddenly, the analysis cannot pinpoint either how long it will last or which agreement will be agreed to in the end. In short, either we lack any theory of rational conflict (if we stick to neoclassicism's equilibrium approach) or we end up, again, with radical indeterminacy and *anything goes*.

3.5 Epilogue: indeterminacy is a prerequisite to *any* rationalisation of conflict

Conflict is the neoclassical economist's greatest challenge. It is, alas, a challenge that they cannot rise to. It is, of course, not for the want of technical sophistication that they fail. The problem runs deeper and touches upon the nature of the phenomenon and the impossibility of rationalising it by means of the neoclassical analytical method. Put bluntly, an economics that insists on analysing the social world in terms of equilibrium behaviour will never be able to shed light on why rational people may, nevertheless, occasionally allow their disagreements to spill over into mutual damage, loss or even carnage. The importance of this theoretical failure cannot be overestimated. For it is a window through which we can peruse the deeper causes behind mainstream economists' 'difficulties' with all types of disequilibrium phenomena; unemployment, unanticipated financial implosions and, generally, the cut and thrust of real capitalism.

In this chapter we asked a simple question: Why do people fight? Is it only mad men and women who resort to conflict? Or, can we imagine perfectly rational agents, who respect deeply each other's rationality, falling into the trap of mutually damaging behaviour? To answer these questions, we deployed the highest form of neoclassical economic theory: the theory of games. Soon we discovered that conflict can be explained easily in a non-cooperative game theoretical context. Just as, in the standard prisoner's dilemma players get locked in a suboptimal equilibrium of mutual defection, so too in various conflict games we can imagine them ending up in a Hobbesian trap,

opting for generalised war because they have a dominant strategy to avoid peaceful behaviour, even when they know perfectly well that peace is preferable to war for each and every one of them.

While non-cooperative interactions can explain conflict, the question then becomes: If players are rational, and can see that the non-cooperative structure of their interaction is leading them inexorably into war's cruel arms, why do they

not agree amongst themselves to hold negotiations the purpose of which would be to culminate in a settlement, a covenant? Why can they not create the institutions that permit binding agreements which offer them an escape from the suboptimality – the horror – of conflict? Note that this is precisely the question that Thomas Hobbes (1991) posed in 1651 in his celebrated *Leviathan*, a treatise that produced the first contractarian theory of the state's legitimacy.

Granted that, as Hobbes had foreshadowed, truly rational agents would want to convert a non-cooperative game into a bargaining process, the question then becomes: Would rational bargainers ever fail to reach an agreement, when there are mutual benefits to be had? Is conflict a possible outcome of rational bargaining? To answer these questions, the chapter turned to the leading authorities of bargaining theory; i.e. of the branch of game theory that analyses bargaining games – John Nash and Ariel Rubinstein in particular. Following the trail that they blazed, we stood by as they stripped the negotiating process to its bare essentials, hoping to uncover the determinants of actual agreements. Unfortunately, we ended up with more than we had bargained for. Instead of analysing and theorising the probability of conflict, the offered analysis eradicated it utterly. The repercussion is that a rational society should also be in a position to eradicate conflict simply by instituting a legal framework which (a) makes contracts inviolable, and (b) allows free communication between antagonists (recall the Gul and Sonnenschein theorem). If it were not for the suspicion that limitless peace is the result of suspect analytical method, rather than a triumph of Reason, this would be very good news indeed.

Before pinpointing the precise fallacy at the heart of game theory's (and neoclassicism's more generally) analysis of conflict, this may be a good point to remind the reader of neoclassicism's penchant for emulating early physics. The analysis begins with frictionless models of some system and derives its basic theorems or 'laws.' Then, friction is gradually introduced into the analysis and the theory's formulae become more realistic at the expense of greater mathematical complexity. In the same vein, bargaining theory at first assumes perfect information between negotiators and shows that, as long as they are fully informed, they will reach an agreement conflict-less-ly. Then, uncertainty is introduced. However, because of analytical difficulties, it is an odd kind of uncertainty that sees the light of day, as it comes complete with the presumption that all eventualities, every possible bargaining tactic, are known in advance and that each comes with a probabilistic belief attached to it. In a sense, it is as if Jill knows fully Jack's complete 'population' of potential selves, together with the likelihood of each, without knowing which of these selves is in Jack's 'driving seat.'

On the basis of this eccentric type of uncertainty the analysis shows that conflict is possible, as a signalling device, but only when exogenous restrictions are in place on the speed with which bargainers can revise their offers and issue new demands. The moment that these restrictions are removed, the probability of costly delay in reaching the uniquely rational agreement vanishes. Hence, the transition from a 'frictionless' to a 'friction'-preserving model does not affect significantly

the dominance of peace over conflict, although it does redistribute benefits or value from the less to the better informed party. Conflict may not materialise but the outcome is determined by the 'power' of agents. While this makes intuitive sense, what exactly do neoclassicists mean by 'power'?

Just as uncertainty is given a narrow meaning, so is power. In bargaining and game theory, power boils down to a combination of relative risk aversion, of relative patience and of the degree to which Jill is better informed about Jack's disposition (relative to Jack's information over Jill's). The real power, the power to change the rules of the game or, more importantly, to impose a convention for its resolution, is ignored. Which brings me to my main concern with neoclassical analyses of bargaining and conflict: It is, I fear, impossible to peel off uncertainty from the bargaining process completely, even if we assume that each side has absolute knowledge of the motivation and capabilities of the other side. Since perfect knowledge is not the same as omniscience or, indeed, telepathy, rational bargainers will always be able to engineer doubt in the mind of their opponents simply by selecting behavioural patterns that the other has not fully anticipated. If this is so, and I believe it is (see below as well as the next chapter), then the very method chosen by bargaining theory's greats is fundamentally flawed: their attempt to analyse at first a bargaining process in which players are fully transparent, only to introduce a degree of opacity later, is thwarted by the fact that rational agents (unlike projectiles or electrons) can always choose to subvert the rules that 'ought to' govern their behaviour. In short, they are and can never be transparent. Not even in theory. Since they can choose to subvert the rules of rational behaviour, there will always exist a significant probability that their opponent will be bamboozled by this 'subversive' behaviour, and, as a result, become more pliable and less demanding. If so, subversive behaviour may well prove profitable (and thus perfectly rational) and stepping off their equilibrium behavioural path may be the quickest road to maximum private gains. But then bargaining theory's model of 'rational' offers and demands will, by definition, fail to delineate the complete set of rational bargaining strategies.

To see this objection to the neoclassical method a little more clearly, suppose that there exists an ultimate theory of bargaining *T* which analyses *all* the relevant information and works out *the* optimal strategy for each bargainer. Assuming that Jill and Jack are rational, and share common knowledge of this fact, each expects the other to choose the strategy proposed by *T* with certainty. What if, however, Jill chooses a strategy other than that which *T* instructs her to choose? Is there no chance that Jack will interpret this deviation as a sign that she is less rational than he had imagined? And if this thought makes him more likely to make concessions to Jill (that *T* does not instruct him to make) since he now fears conflict more, courtesy of the thought that Jill may be more reckless than originally anticipated, why is it irrational of Jill to contemplate violating the instructions of theory *T*? Consequently, if it is not utterly foolish of Jill to consider bargaining strategies (possibly including outright even if limited conflict, that theory *T* has ruled out), then it is clear that theory *T cannot* be the ultimate theory of rational bargaining. And since this is not specific to theory *T*, the contradiction just reached leads to the

safe conclusion that no ultimate theory *T* can *ever* exist. And if no uniquely sound theory of conflict can exist, we are forced to accept that bargaining is profoundly indeterminate.

14

And here is the rub: Either we embrace indeterminacy in any theory of bargaining and conflict (effectively confessing to the impossibility of a determinate theory) or we try to impose upon bargainers a type of equilibrium behaviour which, on the one hand, eradicates conflict but, on the other, is indefensible on the grounds of ... rationality. Yet again, a remarkable and aesthetically beautiful neoclassical research programme has hit the Wall of Indeterminacy. And how has the profession responded? With its usual recoiling into the bosom of wholesale denial, of course.

VERDICT: Neoclassical economists, especially the more philosophically minded, always saw a theory of conflict as their holy grail. They valiantly struggled, for decades, to produce such a theory in a bid to tell a good tale of why people, firms, nations engage

in mutually harmful conflict. After Nash (1950) and Rubinstein (1982), they thought that they had 'nailed' it. They thought that, simply by adding a little uncertainty into their models' perfect information versions, they would be able to rationalise conflict as what happens when rational agents attempt to signal to one another something about themselves: what their true bargaining and conflict costs are, which disposition they have ended up with.

So, a theoretical challenge was issued endogenously, within neoclassicism, to produce a theory of conflict. That challenge was taken up enthusiastically. Yet again, however, this ambitious research project crashed into the Wall of Indeterminacy. The result was that neoclassicism had to choose between (a) an honest admission that bargaining and conflict are indeterminate, and can only be so because rational humans may profitably subvert all equilibrium narratives on how they ought to act, and (b) a dishonest attempt to keep on pretending that their equilibrium analysis maintains a monopoly over rational bargaining behaviour. Naturally, they opted for the latter. To stake this claim, they had to impose an ironclad version of neoclassicism's third meta-axiom (see

[Chapter 1](#)

) which is tantamount to a hidden axiom that non-equilibrium behaviour is ruled out.

In so doing, the profession took path **b** (the backslide) in

[Chapter 1](#)

's diagram in

[Figure 1.1](#)

. Thus, in terms of the *dance of the meta-axioms*, it is fair to say that the theoretical failure to model conflict has produced a **1→2→3→4 shuffle** (see

[Section 1.3.3](#)

). And how has this theoretical failure reinforced neoclassicism's discursive power? Simple. The reader will have noticed that the preceding analysis is immensely complex (even though I have simplified it as much as I could). No person unschooled in its intricacies can even recognise the sleight of hand that is necessary to keep these models together (recall Rubinstein's hidden assumption). This ocean of mathematical complexity acts as a great deterrent for all other social scientists who may have a bone to pick with neoclassicism. As for young graduates, who spent countless months and years mastering this analysis, they will, indeed, require an heroic disposition to 'come clean'; to admit that all this investment has led them to the conclusion that conflict is ... indeterminate. Those

of them who do say this courageously will never get tenure, as their papers will remain unpublished – their models will not have achieved the requisite 'closure'. And those who maintain silence of the faulty foundations of their analysis, continuing to produce models of greater complexity along the same lines (and on the basis of the same denial of indeterminacy's actual hold), will become the new blood that keeps neoclassicism fresh and forever dominant within the economics departments of the best universities.

Notes

1

This chapter draws not only from my *Rational Conflict* (Varoufakis 1991) but also from Varoufakis (1990b, 1992) as well as from

[Chapter 4](#)

of Hargreaves-Heap and Varoufakis (2004).

2

To check that this is so, consider A's intention to choose $X(>0)$ on the belief that someone else in this group, call her B, will select $2X$. For this to be so, A must entertain the expectation that B thinks *mistakenly* that there is someone else, say C, who will select $4X$. Thus any decision to choose a number greater than zero is predicated on beliefs centred upon the prediction that someone (B in this case) is acting on false predictions. In conclusion, the *only* action which does *not* need to be founded on the assumption that someone along the line will hold mistaken beliefs, is the action of selecting zero. Nash therefore 'solved' games by discarding all beliefs which lead to actions which, in

turn, contradict the beliefs which brought them about.

- 3 For a detailed exposition of Nash's equilibrium concept, see Hargreaves-Heap and Varoufakis (2004), [Chapter 2](#). For a reader-friendly proof of Nash's theorem see Dutta (1999).

- 4 An asset can be exogenous (e.g. a windfall profit or an inheritance) or, indeed, one that may be due to *their* actions, past or future (e.g. the result of some partnership, the non-labour surplus of a firm, the gains from trade in the context of a bilateral monopoly or multilateral agreements as in the case of the world Trade Organisation).

- 5 Of course, as we saw repeatedly in previous sections, this does not mean that agents achieve maximal utility. Very often they undermine one another (recall the *Prisoner's Dilemma* and the *Race to Zero*) and, hence, their utility suffers because they are pursuing it so ruthlessly!

- 6 For example, Jill may value the last 1% of her overall share of V in inverse proportion to her overall share of V , whereas this may not be so for Jack. Or, equivalently, once she has (say) secured a certain portion of V , Jill may fear risking disagreement, by demanding even more, more than Jack does.

- 7 Let Jill's and Jack's utility functions be $u(x_L)$ and $v(x_K)$ respectively. Nash predicts an agreement (x_L^*, x_K^*) such that $100 - x_L^* = x_K^*$ and $u'(x_L^*)/u(x_L^*) = v'(x_K^*)/v(x_K^*)$. The last equation says that, at the agreement, the ratio of Jill's marginal utility from her share of the 'pie' to her utility from that share, will equal Jack's ratio of marginal utility from his share of the 'pie' to his utility from that share. Simple manipulation of that equation leads also to the conclusion that the proposed agreement maximises the product of their utilities $u(x_L) \times v(x_K)$. *Proof:* Since $u'(x_L^*)/u(x_L^*) = v'(100 - x_L^*)/v(100 - x_L^*) \Rightarrow u'(x_L) \times v(100 - x_L) = u(x_L) \times v'(100 - x_L)$. But this last equality is the first-order condition for the maximisation of the product of utilities $u(x_L) \times v(x_K)$. *QED.*

- 8 Note that the following is not Nash's own proof. My proof is based on a narrative which, while analytically equivalent to Nash's, brings out the behavioural aspects of Nash's bargaining solution (for its full version see [Chapter 4](#) of Hargreaves-Heap and Varoufakis, 2004).

- 9 To see this, recall that acceptance of Jill's 79.9 per cent offer means that Jack can receive \$79.9 here and now. Now suppose he rejects that offer, insisting that he should get 80 per cent of the \$100 (i.e. \$80). If Jill acquiesces, his payoff will equal 80 per cent of $(100 - 0.6M)$, where M is the time Jill takes (in minutes) to accept Jack's terms. It is easy to see that if $M > 0.21$ (i.e. 12.5 seconds) Jack is better off accepting the 79.9 per cent offer than holding out for an 80 per cent share (even if Jill is expected to capitulate with certainty).

- 10 Suppose that an offer is rejected or accepted instantly. If it is rejected, a counter-offer is issued (again instantaneously) by the rejecting party. However, once a counter-offer is made, there is a fixed time, say ten minutes, the other party replies to this counteroffer. And so on. It is easy to see that this exogenous delay gives the player who kicks off the negotiations a significant strategic (first-mover) advantage: If Jill offers first, a rejection and a counter-offer by Jack will delay agreement by at least ten minutes. She also knows that Jack knows that. In other words, she begins the negotiations under the common knowledge that her opponent can only reject her opening offer by taking a fixed slice of the overall pie (as we have also assumed that with every second that passes without agreement, the pie shrinks by a fixed percentage). Thus, Jill enjoys a *de facto* strategic advantage over Jack, courtesy of her opportunity to issue an offer before the clock starts ticking; an opportunity which means that Jack faces a fixed cost in rejecting Jill's offer. Now, if the minimum response time (i.e. the cost of delay, or the rate of the pie's shrinkage) is less than ten minutes, Jack's fixed cost of rejecting Jill's opening offer diminishes. As the minimum response time (or the rate at which the pie shrinks) tends to zero, Jill's strategic, first-mover advantage vanishes.

- 11 For example, if $\alpha = \beta = 0.8$, two rejections will mean the loss of 36% of the pie; while failed offers will 'destroy' more than 50% of the value that was initially available for distribution between the two bargainers.

- 12 Before discussing this noteworthy result further, we note that equality $1 - \beta(1 - \alpha V) = V$, on which Rubinstein's solution rests, demands that our bargainers have the same discount rates at $t = 1$ as they would later on in the game.

- 13 Note that this subversive plan on Jack's behalf is analytically equivalent to R's subversive thoughts in Section 3.5.1.

- 14 The reader will note that this conundrum constitutes a mirror image of the paradox of rational conflict with which this chapter began.

4 No bluffing please, we are economists!

Why bluffs and other subversive acts preclude determinate game theoretical analyses

4.1 Prologue

4.1.1 Background briefing

Economists like to think of themselves as the purveyors of the best analyses of rational human behaviour. They are quite happy to let psychologists study the thoughts and acts of the irrational, of the clinically depressed, of the delusional, as long as their monopoly on rational behaviour is recognised. When others point out that most people harbour inconsistent preferences, act in a manner that is frequently at odds with their own perception of self-interest, entertain ridiculous beliefs etc., economists nod approvingly but then immediately retort that their subject matter is the 'ideal' rational self. And that what makes their models of this über-rational *homo economicus* relevant is the 'fact' that markets have a tendency, through some Darwinian process, of eliminating behaviours that diverge from the rational choice of a *homo economicus*. In their mindset, the greater one's exposure to market competition, the more one begins to resemble the ideal type rational agent that they, and their models, take for granted. Even if most of us fall short of *homo economicus*' capacity for rational choice, claim the economists, we cannot help but tend to that ideal type the greater our exposure to the cut and thrust of market societies.

In this sense, economists do not mind it when other social scientists disparage their model of men and women as unrealistic, as unrepresentative of how people actually think and act, even as downright misleading about the people around us. Their defence is simple: while *homo economicus*, the instrumentally hyper-rational ideal type, may not exist, it is an excellent benchmark against which to 'measure' the rationality of living and breathing humans and, moreover, it represents a very helpful model of the type of behaviour toward which real humans tend the greater their immersion in market competition. And when critics of the economists' theories point out systematic differences between actual behaviour (e.g. in the laboratory) and the behaviour economists predict, the latter resort to the explanation that these differences are the result of the fact that people are not as rational as they, the economists, assume. That if they were *truly* rational, economic theory would predict perfectly their behaviour. Thus, economists interpret the chasm between observed human behaviour and the behaviour their

models predict as a reflection of the divergence between actual and ideal human rationality.

The previous chapter should have already alerted the reader to a serious objection to this claim. For we have seen that neoclassical economics' theories of conflict between rational people fail, not because humans are less rational than the economists assume but, indeed, because they are *more* rational than that. If this is so, and I have no doubt that it is, the economists' claim of a monopoly of rational strategic thinking is bunk and, to boot, their excuses for the predictive power of their models is utterly unconvincing.

¹
The point of substance raised in
[Chapter 2](#)

is that economic models of conflict fail *by design* for the simple reason that, if they were good and proper depictions of rational strategic behaviour, then rational agents should be able to use them in order to predict what their opponents will do and, also, to predict what their opponents will think that *they* will do. But, if this is so, then truly rational players should be able to ask a deeply subversive question: 'What if I violate the

rules which, according to the best model, ought to govern my behaviour? Is there no chance that my opponent will be confused? And if so, might I not be able to extract benefits from her confusion?' If the answer to this question cannot be assumed to be reliably negative, then all of a sudden the model in question will no longer be able reliably to predict what rational people will do. Its failure will not be due to the agents' less than perfect rationality but, remarkably, it is brought on by the fact that they are so rational as to attempt to use the model to their advantage, so much so that they destroy its predictive powers!

These thoughts lead to a simple conclusion: If we are to acknowledge the true powers of human reasoning, we must admit that even the most brilliant model of their strategic behaviour will fail to pinpoint the full set of optimal choices. So, either we respect human rationality to the full, but acknowledge that humans' rational behaviour is radically indeterminate, or we continue to search for the determinate model which, nonetheless, will only hold analytical water if we refuse to acknowledge the rational agent's capacity to profit from subverting the model's predictions.

In short, the price of determinacy is the denigration of human rationality; a trade-off that is precisely contrary to the economists' claim on behalf of their models. While economists claim that their models fail only to the extent that they assume too much rationality, in reality they fail because they assume too little rationality on behalf of agents.

4.1.2 The rest of this chapter

In 1993 I published an article in *Erkenntnis*, an analytical philosophy journal, in which I took further the theme above. It was entitled 'Modern and Postmodern Challenges to Game Theory' and aimed at driving home the point that economic models of strategic interaction do two things: First, they presume a great deal of computational power on behalf of human agents. Secondly, that they, at the same time, deny agents possessing such super-computing power the 'right' to even

imagine that they can bluff successfully against similarly hyper-rational opponents. The central question is: Is this because game theory proves that any such attempt at bluffing will necessarily fail? What I showed in that paper, which is reproduced below with several emendations, is that game theory can *never* demonstrate that bluffs are doomed to fail. Thus, it cannot prove that rational agents should *not* consider bluffing. The only reason that game theory, and economics more generally, introduces (through the back door) the hidden assumption that bluffs do not pay (and that rational agents dismiss them as strategy options) is because otherwise their models would remain indeterminate. Faced with a choice between intellectual honesty and closing their models, economists never fail to opt for the latter.

The next section positions game theory as the highest form of neoclassicism and as an important, yet narrow, part of the overall Enlightenment project.

[Section 4.3](#)

introduces the reader to the manner in which game theorists come to conclusions on how rational agents reason, while

[Section 4.4](#)

challenges this view. Then

[Section 4.5](#)

gives economists their chance to respond to this challenge with an analytically brilliant answer which, nonetheless, is exposed as logically incoherent in

[Section 4.6](#)

.

[Section 4.7](#)

generalises ambitiously and uses this debate (between game theorists and critics like myself) in order to draw some broad conclusions about the Enlightenment project

and the way it has been affected by the dominance of neoclassical economists. Finally, Section 4.8 concludes and links these debates to the book's overall theme regarding the irrepressible emergence of indeterminacy every time economists try to 'close' their models.

4.2 Game theory and the Enlightenment project

The battlelines in social theory have frequently been drawn along two familiar views of human agency. First, there is the perception of the sovereign agent whose autonomous desires forge the social structures that will fulfil them. Society is, therefore, seen as the means by which the agents' ends will be instrumentally realised. Secondly, there is the view of an individual whose desires are the product of social structure. Even if the agent rationally pursues her objectives, she is still a plaything of social forces which she cannot control.

In this chapter the above controversy is bypassed in favour of a deeper controversy between the dominant variant of modernity and its foes. To accomplish this, I focus on a well-known game in which agents have given payoffs and a unique equilibrium strategy. That a fierce controversy is engendered in a simple framework is evidence that one does not need complex social interactions in order to end up with complex social phenomena. That this controversy also has the potential of inciting clashes between Humeans, postmodernists and Hegelians, is an indication that game theory ought to be more than a search for clever strategies. Indeed, as I will argue, it should be a tool for exploring the meaning of rationality in social settings.

The following analysis is based on one particular game (often referred to as the centipede game) although it is not too difficult to show that the main problem

is pervasive in game theory. Sugden (1991) demonstrates the generality of similar concerns. The following pages re-evaluate an increasingly popular critique of game theory's method by constructing a sophisticated defence of game theory only to show that it is ineffective. The ensuing discussion sharpens the critique and allows us to draw parallels between the debate on game theory and some crucial philosophical controversies. In this vein, the analysis is aimed at new interpretation rather than at new solution concepts. Part of the offered interpretation is directed at game theory itself. To give a flavour of what follows, I will propose that we interpret conventional game theory as an extremist faction of the modernity project. Defining modernity as the optimism generated by the Enlightenment concerning the ability of Reason objectively to answer complex questions concerning nature and society, I will conclude that there are three alternatives: (a) to remain within modernity while renouncing its more extreme (neoclassical equilibrium theoretical) branch, (b) to reject modernity's concepts altogether (the postmodern suggestion), and (c) to turn away from the dominant aspect of modernity towards a hitherto neglected version of it.

4.3 Inducing equilibrium beliefs

Suppose we have two individuals whom we pit against each other. We promise them a large sum of money and ask them to find some way of splitting it between them. However, we shall let them collect their reward only if they strike a deal. Furthermore, as the seconds tick away without an agreement, we continually reduce the sum in order to give them an incentive to agree quickly. Can we have a theory of what will happen?

Game theory produces a narrative of what will happen starting with the simplest of cases. Suppose, we are urged, that the two are identical and that they know it. Not only do they share the same objectives, but, also, they are transparent so that each knows exactly what the other desires. If this is the case, game theorists assume that there is a unique outcome provided each agent is entirely rational,

² knows that the other is entirely rational, knows that the other knows this etc. (from now on this assumption of *common knowledge rationality* will be referred to as **CKR**):

they will instantly settle for a 50–50 split.

Once this result is obtained, game theory relaxes its assumptions progressively and tackles more complex versions of the same problem. First, it allows for differences in attitudes toward risk and rewards the (relative) risk-taker with a greater payoff, and, secondly, it introduces asymmetric information in order to show that the possession of more information is advantageous (see Harsanyi 1973) – see the previous chapter for a detailed presentation of this type of analysis.

Why is the above example representative of contemporary modernity? Two individuals facing each other in an instance of pure antagonism develop trains of thought which swiftly terminate the cacophony that would have arisen in a pre-modern narration of their situation. Instead of the drama of equally intelligent belligerents duelling to the last for personal gain, Reason is called upon to furnish Harmony and Efficacy. The fact that their rationality is common knowledge

is presented as the bedrock of a uniquely rational train of thought that each will, if rational, latch on to. The clear separation of Reason from Unreason allows the theorist to view agents as identical computers running the same software with identical initial information (input) guaranteed by the assumption of perfect information. It is not surprising that they will inevitably come to the same conclusion (output) and thus agree without delay.

An immediate postmodern concern is that the computer metaphor inhibits understanding of human responses since our reasoning can be neither unique nor transparent. A more radical postmodern objection is that to talk of Reason is to talk of a term that has no concrete equivalent in social reality. It is not that agents are irrational, but that it is unclear what it means to be rational in social interactions. If this is so, the analogy with the numerical algorithm is misleading.

Whereas game theory treats simple social interactions with given objectives in a way that the outcome can be assumed and used later for explaining more complex situations (the *analytic-synthetic* road to explanation), postmodernity claims that the only chance of defensible choices, even in simple situations, materialises when we recognise the impossibility of understanding our reasoning by means of metaphors that devalue and oversimplify.

In a manner reminiscent of Parmenides' definition of nothingness (i.e. a radical absence of reality

) game theory identifies Reason with the residual left behind once Unreason has been expelled. By contrast, postmodernity claims that human thoughts are irreducible to a field where Reason is dominant. Critics of the type of extreme modernity that lies behind game theory, and of course mainstream economics, have frequently accused it of having a social-less theory of individual agency, of procuring a process without a subject. This is not the postmodern position. The latter denies both the possibility of subjectivity *and* of analytically breaking down complex social interactions into simple ones before synthesising the resulting insights into a general social theory.

It is interesting to explore the connection between the postmodern critique which has been developed by writers on literary criticism and philosophy (and who have probably never considered game theory) and recent criticisms developed by game theorists themselves.

Consider the interaction between A and B in

Figure 4.1

;

Potentially, there are three stages in this game which begins with A having a choice

of putting an end to it (by playing UP) or passing on the baton to B (by playing DOWN). If A chooses the latter, it is up to B to choose whether the game will proceed to $t = 3$. If it does, then A has the final say. Glancing at the payoffs, two things become clear. Both players are better off if $t = 3$ is reached than if A terminates the game at $t = 1$. On the other hand, A can see that if the game is to end at $t = 2$, rather than at $t = 3$, she would be better off putting an end to it right at the outset. Supposing that they have no way of communicating with each other either prior to the play of the game, or during it, other than through their UP/DOWN choices, is there a way of predicting with certainty what they will do?

Before answering, game theory introduces its axiom of *common knowledge rationality* **CKR**: A knows that B is rational, B knows that A is rational, A knows that B knows that A knows ... that B knows ... that A is rational – *ad infinitum*.

And what does 'rational' mean? It means that, if there is a strategy which maximises one's payoffs, one will recognise it and adopt it. So, a rational A must try to work out whether it is better to play DOWN during $t = 1$, thus giving B the option of ending or continuing the game, or to play UP, collect payoff 1 and end it there and then.

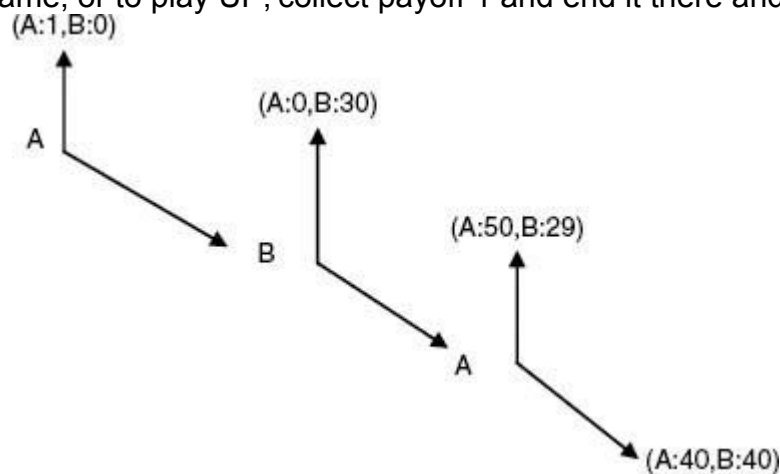


Figure 4.1

The centipede game.

Her decision hinges on what she expects B to do at $t = 2$. If she is convinced that B will choose UP, then she ought to give him no opportunity of doing so, since her payoff would be 0 compared to the 1 from playing UP at $t = 1$. If, on the other hand, she expects him to play DOWN, then she should let him do this because reaching $t = 3$ will endow her with payoff 50. However, game theory claims that this is an expectation she will never entertain.

This conclusion is reached as a result of A's following thought process, while she is attempting to predict B's thoughts at $t = 2$, were he to be given a chance to get to that stage. Player A tries to get into his shoes and imagine what thoughts she would entertain at $t = 2$ and, therefore, what decision she would reach (UP or DOWN). A thinks:

B will play UP at $t = 2$ if he expects that his payoff from doing so, i.e. 30, is greater than what he can rationally anticipate at $t = 3$. At $t = 3$ I am the one who does the choosing and I will clearly play UP leaving B with payoff 29. Since this is less than what he will get from ending the game at $t = 2$, it is silly of me to expect him to play anything other than UP. Thus, the conclusion that $t = 3$ will not be reached leads me to the conviction that I am better off by playing UP at $t = 1$.

The above logic is based on backward induction and generates the unique equilibrium set of beliefs that allows A to come to a conclusion about the best course of action.

Figure 4.2

shows, it unfolds backwards, beginning with a conjecture at $t = 3$ which leads to A's final conjecture at $t = 1$. The process that takes A from (c) to (e) is underpinned by **CKR**, i.e. (a) and (b):

Assumptions:

- (a) AbB is rational
- (b) $AbBbA$ is rational

Fundamental conjectures:

- (c) $t=3$ C1: $A \rightarrow UP$
- (d) $t=2$ C2: $BbC1$ thus $B \rightarrow UP$
- (e) $t=1$ C3: $AbC2$ thus $A \rightarrow UP$

A's composite conjecture inducing the equilibrium outcome:

- (f) $AbBbA \rightarrow UP$ at $t=3 \Rightarrow A \rightarrow UP$ at $t=1$

where b and \rightarrow denote the verbs *believes* and *plays* respectively.

Figure 4.2

The logic of equilibrium strategy ($A \rightarrow UP$ at $t = 1$).

Earlier I referred to the game theoretic predilection to the assumption that two agents with identical payoffs, rationality and information are, ontologically, identical. They are to be seen as 'running' on identical algorithms or software and coming to the same conclusion. If this is correct, then A can replicate B's thoughts perfectly since she can put herself in his shoes by pretending that his payoffs are her own. This is what allows game theorists to assume that the division game described earlier is trivial and also that the passage from (c) to (e) and (f) in

Figure 4.2

ought to be automatically accepted.

To sum up,

Figure 4.2

is a good example of how the dominant modernity lurking behind game theory analyses a simple interaction between two agents, of how it breaks beliefs down to their elemental components, and uses induction in order to put back together a string of conjectures leading to an equilibrium result.

4.4 The challenge

Game theory establishes a rationalist vision of order which promises to 'solve' complex social interactions. Effectively, it turns social phenomena into the subject of natural scientific discourse. The logic of backward induction in

Figure 4.2

is a simple example of this. It begins with assumptions concerning the rationality of agents and derives their unique thought process.

Figure 4.3

converts this logic into a computer algorithm which agents must follow, at least according to mainstream game theory, if they are to qualify as rational reasoners.

The point to note here is that **CKR** renders the above programme common property. It is presented as the uniquely rational sequence of conjectures that one

must have when one seeks to maximise one's payoffs. Agents are assumed to recognise in it the optimal algorithm before they are allocated the role of A or B. Thus, they have worked out its logic in advance and expect that a rational A will play UP

during the first stage if the payoffs are as in

Figure 4.1

- STEP 1 Compute P_3 as your maximum payoff at $t=3$ in the following manner: if you are player A, choose P_3 as the largest payoff; if you are player B, choose P_3 as the payoff you will collect when player A choose her largest payoff
- STEP 2 Compute P_2 as your payoff at $t=2$ if the game is ended here
- STEP 3 If you are player A got to STEP 6; otherwise continue
- STEP 4 If $P_2 < P_3$ play DOWN at $t=2$; if $P_2 > P_3$ play UP at $t=2$; otherwise randomise with equal probabilities at $t=2$
- STEP 5 STOP
- STEP 6 Compute P_2 as your payoff at $t=1$ if the game is ended here
- STEP 7 Play DOWN at $t=1$ if either (a) at STEP 4 the decision is to play DOWN and $P_1 < P_3$, or (b) at STEP 4 the decision is to play UP and $P_1 < P_2$. Otherwise play UP.

Figure 4.3

The strategy of rational agents according to

Figure 4.2

Not surprisingly, when roles are finally assigned, whoever gets the part of A plays UP instantly. **CKR** does for equilibrium game theory's view of Reason what the veil of ignorance does for John Rawls' (1971) concept of justice: it defines it *via* a process of de-personalisation. The second point to note is that the adoption of this program requires that Reason is a means by which agents (as well as game theorists) convert an expectation into a conviction. For example, at stage $t = 1$ player A is facing a choice between a certain reward (payoff 1 if she plays UP) and a conjecture concerning what she will end up with if she plays DOWN. Backward induction, faithfully reproduced in

Figures 4.2

and

4.3

, turns this conjecture into an absolute conviction that, were she to play DOWN, her payoff would be 0.

It is now time to explain why the above is highly problematic. The critique of backward induction which follows has been around for some time

9

but its impact has not been felt outside the narrow circles that produced it. Nevertheless, it is an important critique with repercussions for the way social theorists incorporate game theory in their models but also because it allows us to place the debates between game theorists within the larger debates in social theory. It begins with a devious thought that may cross A's mind:

I understand

Figure 4.2

well and I agree with assumptions (a) and (b). Therefore, I see why its logic should lead me to the conclusion that UP at $t = 1$ is the only sensible strategy for me. However, what if I choose to defy it?

For A rationally to pursue this thought, she must be able to support it by a consistent train of conjectures similar in structure to those in

Figure 4.2

Figure 4.4

presents such a sequence. The question is whether it is rational to entertain such conjectures.

Game theory's conventional response is that thought processes like the one in

Figure 4.4

are incompatible with rationality. The deviant logic in

[Figure 4.4](#)

is axiomatically ruled out on the basis of **CKR**. If agents take the

[Figure 4.3](#)

algorithm to be the best way of playing the game, then the subjective probabilities p and q in

[Figure 4.4](#)

must be zero at all points in logical time. If this is so, a rational A who is linked mentally via **CKR** to a rational B will *never* contemplate any strategy other than UP at $t = 1$. But why should players believe that the

[Figure 4.3](#)

algorithm is the one they ought to follow?

The crucial point here is that, if **CKR** is a necessary condition for dominance of

[Figure 4.3](#)

, but is a condition that rationality itself cannot support, then there may be an opening for

[Figure 4.4](#)

. Let us define a deviant choice as one which goes against the equilibrium prescription of game theory but which may or may not be irrational. For example, in our game (see

[Figure 4.1](#)

), if A ever played DOWN at $t = 3$, we conclude that A is (instrumentally) irrational. However, if A plays DOWN at $t = 1$, then she may or may not be irrational depending on her ability to justify her choice in terms of her objectives and a set of rationalisable beliefs. If she can justify her belief in the superiority of playing DOWN as a strategy for reaching the (50, 29) outcome, then her strategy is deviant albeit not irrational.

The logic of

[Figure 4.4](#)

can be summed up thus: At $t = 1$ player A contemplates playing DOWN instead of her equilibrium strategy UP for a simple reason: she is hoping that by so doing $t = 3$ will be reached. Why? She thinks to herself:

If B is convinced that at $t = 3$ I will play UP then he will always play UP at $t = 2$ and then we will never reach $t = 3$. Thus, if I believe that this is what

he thinks, then I should choose my equilibrium strategy and play UP at $t = 1$. Indeed, according to (a) and (b) in

[Figure 4.4](#)

, I know that he believes most strongly that I am rational and, therefore, he currently expects with probability 1 that, in the hypothetical case that we reach $t = 3$, I will play UP. So, at first glance I should conform with the equilibrium logic of

[Figure 4.2](#)

. However, according to (b) in

[Figure 4.4](#)

, this is exactly what he expects me to do. What if I do not oblige and play DOWN at $t = 1$? Surely, he must sit back and take notice.

Assumptions:

- (a) $A \rightarrow B$ is rational with probability $1 - w = 1$.
 $B \rightarrow A$ is rational with probability $1 - p = 1$.
- (b) A and B know (a).
- (c) If at $t=2$, p were to equal 1 then: $B \rightarrow A \rightarrow \text{DOWN}$ at $t=3$.

Definitions:

- (d) Let $p' > 0$ be the probability belief of B at time $t=2$ that would induce B to play DOWN at $t=2$.
- (e) Let q be A's probability belief that p exceeds p' ; i.e. $q = \Pr(p > p')$.
- (f) Let $q' > 0$ be the probability belief of A at time $t=1$ that would induce A to play DOWN at $t=1$.

Fundamental conjectures:

- (g) A believes that if she defies the logic of backward induction and plays DOWN at $t=1$, then B will revise p upwards at $t=2$.
- (h) $q > q'$ at $t=1$ and, therefore, $A \rightarrow \text{DOWN}$ at $t=1$.

where b and \rightarrow denote the verbs *believes* and *plays* respectively.

Figure 4.4

The logic of the deviant strategy ($A \rightarrow \text{DOWN}$ at $t=1$).

This last thought is the gateway to the deviant logic of

Figure 4.4

. In trying to anticipate what B will think, agents are forced to stop operating like automata, to ditch the program in

Figure 4.3

, and to start *thinking* as opposed to following formulae.

10

In this context, player A may continue her reflection thus:

Since my choice deviates from that of the

Figure 4.2

recommendations, he will be forced to find an explanation. There are two possibilities. One is that he will think that I am irrational for not doing as

Figure 4.2

prescribes. If this is so, he will change his game plan and play DOWN at $t=2$ expecting my irrationality to overcome my senses so that at $t=3$ I will choose DOWN. Of course, there is the other possibility that I must reckon with. Player B may realise that this is exactly what I am thinking and refuse to believe that I am irrational simply because I have chosen irrationally. Nevertheless, all I need in order to consider playing DOWN is that B assigns a relatively low probability that I am irrational; not that he is convinced of my irrationality. Let p be the non-zero probability that he assigns to this prospect after observing my deviant choice at $t=2$. If $p > 1/11$ (in terms of part (d) in

Figure 4.4

, $p' = 1/11$), then his expected return at $t=2$ from playing DOWN exceeds that from UP, therefore giving him a strong incentive to deviate from his equilibrium strategy too, i.e. play DOWN at $t=2$. So, I conclude that if my defiance of the logic of

Figure 4.2

makes him think with probability $1/11$ that I am irrational then it may, after all, make sense for me to play DOWN at $t=1$ since there is now a realistic chance of getting 50 at $t=3$ rather than 1 at $t=1$.

We have come full circle. Player A accepts the assumption that B believes her to be rational with probability 1 but is prepared creatively to explore the thought that deviant

behaviour must make those who *ex ante* rule out the possibility that their opponent is irrational to suspect *ex post* that, after all, she may be irrational. She concludes that following her explicitly deviant behaviour, if B's *ex post* belief in her irrationality becomes positive (1/11 in our example), it may make sense to behave in a way that game theorists would consider irrational. More precisely, if A expects p to exceed 1/11 with probability a touch over 1/50 – i.e. if A expects that there is a 1/50 probability that her deviance at $t = 1$ will make B think that she is irrational with probability 1/11 – then her expected returns from playing DOWN in defiance of game theory's logic are greatest. Hence, part (g) in

Figure 4.4

.

Figures 4.2

and

4.4

offer alternative logics that A can choose from. Can they be equally valid? Game theorists favour the equilibrium story on the basis that it is uniquely compatible with **CKR**.

12

Under this type of common knowledge, player B will never update p upwards if A chooses her deviant strategy at $t = 1$ and, therefore, player A will never entertain a subjective probability q that exceeds 0. But this is too strong. As Pettit and Sugden (1989) have shown, a subtle difference in how we interpret shared rationality can change all this. All we need is to treat shared rationality as something agents believe in rather than as an immutable axiom. If we assume that agents *believe* that irrationality is absent at all orders of belief, instead of axiomatically dismissing any possibility of doubt concerning the presence of irrationality, then the deviant strategy is given a chance. The difference becomes apparent when we look at part (g) in

Figure 4.4

and compare it with parts (c), (d) and (e) in

Figure 4.2

. In the former case a deviation from what is deemed to be rational behaviour has the potential of making B wonder whether his cast-iron belief in A's rationality is well founded. In the latter case, by contrast, A and B follow the predetermined program in

Figure 4.3

since no deviation from the equilibrium scenario will make them wonder about the correctness of their conceptualisation of the game. Thus,

Figure 4.2

requires that, once rationality is assumed, players do what

Figure 4.3

tells them *regardless of whether their opponents choose in the manner that*

Figure 4.2

predicts.

In summary, the point of contention seems to revolve around the agents' subjective beliefs. Game theory leans on **CKR** in order to rule out any uncertainty about the beliefs of one's opponent. Thus, it reduces the set of optimal strategies to the one in

Figure 4.2

and does not concern itself further with the prospect of rational deviations. If we choose a slightly amended version of common rationality which allows agents to re-think their conviction concerning the absence of irrationality once they observe deviations from the

Figure 4.3

algorithm, then deviance can be shown to be rational. Another way of conceiving

our theoretical dilemma is this: under **CKR** agents are incapable of forming views about what they ought to do in the future if they find themselves at a part of the game-tree that **CKR** would not have allowed. They do not need to do so because **CKR** axiomatically assumes that no such trespassing ought to be considered. But when it is considered, agents may conclude that it is in their interest to abandon the equilibrium path. Indeed, would they not be irrational if they failed to consider all outcomes, including those that **CKR** deems unwise? And if the mere contemplation of these parts of the game-tree renders deviance rational (though not uniquely so), is this not conclusive proof that **CKR** is inappropriate?

Defenders of **CKR** may protest that the above argument suffers from the following defect: If A's deviance at $t = 1$ manages to raise B's estimation of her irrationality, then how does B predict an irrational A's behaviour at $t = 3$? And if B has problems at $t = 2$ in predicting A's behaviour, how can we say that A's deviant strategy is rational at $t = 1$ when she cannot know how B will be thinking at $t = 2$? This is a good point. It proves beyond doubt that our players face risky decisions once **CKR** and the safety of the

[Figure 4.2](#)

logic are abandoned. This is, however, no proof of the irrationality of deviance; it is merely

confirmation that neither the equilibrium nor the deviant strategies are uniquely rational.

In effect, when A contemplates the deviant strategy she is hoping that she can deceive her opponent. Is this rational? The answer must be that it is certainly not irrational. There is nothing in the structure of this game to suggest that an instrumentally rational agent ought to assume that she cannot out-manoeuvre her opponent. By the same token, it is also rational to think that she cannot do this. The problem with game theory and its **CKR** foundation is that it instils in agents' minds the belief that deception can never work. It is unclear what institution or psychological mechanism performs the same role in society.

4.5 One negative and one positive defence of the equilibrium approach

Modernity inspired an extraordinary confidence about our ability scientifically to solve complex natural and social problems. The imposition of **CKR** by game theory may be thus interpreted as an extremely confident attempt to consolidate modernity's spirit in games such as the one in

[Figure 4.1](#)

. The previous section challenged this spirit by encouraging agents to ask questions such as: 'What if I do not do what the theory suggests I ought to?' Of course this is not a question that automata can ask. And since game theory models agents *as if* they were automata such as the one in

[Figure 4.3](#)

, then game theory fails to grasp this important dimension in rational agency.

¹³

In this section I will be presenting two lines of defence for equilibrium game theory. The first is a negative defence in the sense that it refuses seriously to consider the alternative (deviant) strategy advocated in

[Figure 4.4](#)

. The second defence is much more sophisticated: rather than ignoring the possibility of deviant play by player A at $t = 1$, the latter tries to explain it by means of an argument that is internal to game theoretical thinking. In true modernist spirit, both defences rely on the belief that there exists a unique theory describing rational play in this game and, indeed, in *every* game. Where they diverge is that the negative defence does not allow Reason to take more than one form *within* the unique theory whereas

the latter does.

Starting with the so-called *Harsanyi doctrine*, a negative defence would claim that, if Reason is unique and unabridged, and if the two players are *equally* rational then (in the absence of asymmetric information), they must generate identical trains of thought.

¹⁴

Figures 4.2

and

4.3

provide the only thoughts compatible with this requirement. If we are to accept the logic of

Figure 4.4

, the negative defence continues, then we accept the possibility that one of the two players may form expectations which are proved wrong by the play of the game.

¹⁵

But since we assumed that they are equally rational, how can we allow one of them to develop correct expectations while the other does not?

The above defence suggests that if we are to assume identical rationality then we must accept the equilibrium logic. Perhaps, this defence argues, it is not a good idea to make this assumption. Then, of course, it is not game theory that we must

blame for producing a result we do not like but our assumptions. However, I do believe that this argument is untenable. For who is to say that if there are two identically rational agents involved in such an interaction, both of their trains of thought must be proved correct? To be prophetic is not a prerequisite for being rational. If, indeed, there is more than one rational train of thought, our players may form different sets of conjectures where each is just as rational as the other. Quite naturally, one may end up with conjectures that are confirmed by the actual choice of strategies while the other does not.

¹⁶

This is not to say that one is more rational than the other.

Relating the above argument to this chapter's broader theme, it seems as if game theory has a tendency to maintain that Reason is more powerful than it can ever be. The negative defence burdens it with the task of coordinating beliefs and choices when, on its own, it can do no such thing. The moment our players are told that (in the context of

Figure 4.1

) their opponent is rational, they are supposed to know exactly what will happen because the thought that one may try to outwit the other never crosses their mind. If it could be demonstrated that equal rationality has this effect, then the defence would be successful. Unfortunately, what I refer to as the negative defence is based on the assumption that such a pernicious thought will not arise. Why not? Because game theorists believe that if two players are equally rational, then we cannot allow a situation where one of them out-manoeuvres the other. Bluffs, in short, can never succeed if the economist is to have her equilibrium solution.

However for this to be logically viable, it must be shown that rationality commands players who hold their opponents in high regard to abstain from efforts to outwit each other. What boring events world title chess championships would be if this were true! Strategic cowardice, and a total aversion to bluffing, cannot be synonymous with rationality, even if compatible with it.

¹⁷

It seems to me that the crux of the argument is that the negative defence demands that agents cannot distinguish between the following two statements:

- (i) My opponent is rational and thinks I am rational, and
- (ii) There exists only one train of thought that is rational to form in this game.

I cannot see why (i) should necessitate (ii) if agents are equally rational. If it does not, the negative defence fails to meet the challenge of

Figure 4.4

and relies on a perception of Reason which is open to what Hegel wrote in the *Phenomenology*: 'It lives in dread of besmirching the radiance of its inner being through action and existence. In order to preserve the purity of its heart, it flees from contact with actuality and persists in a state of self-willed impotence.' The challenge of

Figure 4.4

is denied simply because it cannot be grasped.

Let us now turn to what I described as a positive defence of game theory. Such a defence ought to attempt to undermine the

Figure 4.4

logic by showing that something very similar to the latter can be constructed if we follow the method that gave rise to

Figure 4.2

. In other words, a sophisticated game theorist would argue that the reason why the deviant strategy sounds plausible is because it has

a perfectly good equilibrium foundation, rather than because equilibrium theory is deficient. Thus, game theory would attempt to assimilate

Figure 4.4

rather than to banish it. To do this it accepts the proposition that there may, after all, be more than one rational train of thought.

Before moving to the positive defence it is useful to look at a possible interpretation of the challenge to the original equilibrium theory, as presented in

Figure 4.4

. The latter urges player A to choose UP at $t = 1$ after looking at $t = 3$ and projecting the decision she would have made at that stage onto player B at $t = 2$ and then back onto herself at $t = 1$. In a sense, player A is asked to 'observe' what she would have done at $t = 3$, induce from that what B will do at $t = 2$ and further induce what she ought to do at $t = 1$. Whether this induction is appropriate or not depends on the projectibility of the conclusions derived from an analysis of stages $t = 2$ and $t = 3$ *in isolation from the rest of the game*, onto stage $t = 1$. Game theory uses the **CKR** assumption in order to ensure that the compartmentalisation of the game into subgames separately to be examined is uniquely legitimate. However, by ignoring the projectibility of conjectures from one subgame onto another it neglects an important aspect of rational induction.

Say player A is about to choose her strategy at $t = 1$. According to backward induction, she looks at $t = 3$ first and thus illuminates her current choice. Game theory identifies the ability to 'induce' in this manner a unique rationale with rationality. But is induction invariably trustworthy? For instance, she may ponder the proposition that, in logical time, all stages of the game precede $t = 1$. By induction, may she conclude that all stages of the game will share that trait? This conclusion would lead her to believe that the game will never start since $t = 1$ cannot eventuate. Taken further, a second level induction, an induction about such inductions, tells A that such inductions are always wrong. Should she now believe that the game has started a long time ago since there can never be a stage not preceded by another stage of the same game? Quite clearly, a blind application of induction does not conduce intelligent thoughts. We need something more before we resolve that a trait characterising one stage of the game is projectible onto another.

18

Some traits command confident expectation of continuance from one stage to another and some do not. We do not expect the trait 'being prior to $t = 1$ in logical time' to carry forward to past moments without end. How do we know whether a trait is

projectible or not? A phenomenon that is immediately noticeable and has recognisable form is potentially projectible by means of induction – e.g. a sunset is projectible from one day to another. Postmodern thinkers attach a great deal of importance to language. They would argue that a trait is projectible if there is a word for it that reveals, rather than hides, its true meaning. Game theory, on the other hand, derives the logic of

Figure 4.2

, and pushes hard for it to be recognised as a uniquely rational logic, on the basis of an induction without establishing the projectibility of the main trait that is being carried from $t = 3$ to $t = 1$. The critique of the equilibrium solution to the game in

Figure 4.1

(that preceded) refuses to accept that the meaning of the word ‘rationality’ is clear enough to sanction unconditionally the kind of induction required for the generation of

Figure 4.2

. Therefore, in circumstances of a truly interactive game, the trait ‘rational choice at $t = i$ ’ [where $i = 3, 2, 1$] is as unprojectible as the trait ‘being prior to $t = 1$ in logical time’ above.

19

The result is that

Figure 4.2

cannot represent the uniquely rational train of thought. This is a familiar postmodern view regarding modernist theories which are accused of mistaking analogies for concepts.

20

In our case, game theory mistakes the consistency which results from analogous behaviour (such as that prescribed by

Figure 4.3

) for rationality.

So, a positive defence of game theoretic orthodoxy must begin with a humble admission that

Figure 4.2

is only an embarkation point and not the destination of the equilibrium narrative. The **CKR** assumption should then be interpreted as an initial assumption to be relaxed soon after the theory has got off the ground. The refined game theorist must admit that there are different ways of conceptualising the game of

Figure 4.1

and that it is unwise to *assume* that a rational player A will *never* play DOWN at $t = 1$. However, the positive defence must insist that, if we are to understand what happens when two equally intelligent players participate in this game, we must utilise the tools of equilibrium analysis even if we reject its earlier conclusion based on quite restrictive rationality assumptions. Kreps *et al.* (1982) offer a good basis on which to build such a defence.

The first leg of a positive defence would be to modify the **CKR** assumption as stated in parts (a) and (b) of

Figure 4.2

. In its stead, it would place the assumption that players may now suspect their opponent to be irrational (see (a) and (b) in

Figure 4.5

). Furthermore, it allows for more than one kind of rationality so that player A may be rational and yet not conceptualise the game according to

Figure 4.2

; let me label the latter the *Reason of Backward Induction* or **RBI**. After making these concessions, the positive defence procures its rationalisation of the deviant

strategy on game theoretical grounds. Central to it is the diversity of logics which a rational player may adopt as well as the possibility of flagrant irrationality. In

Figure 4.5

player B expects player A to be irrational with probability p and is convinced that an irrational A will always choose DOWN at $t = 3$.

21

Also, if he thinks she is rational he does not immediately assume that she will adopt **RBI** and the

Figure 4.2

logic. He expects a rational A to deviate from **RBI** with probability $1 - r$ (see part (h)) due to the adoption of an alternative mode of reasoning.

What kind of reasoning is that? It depends on how open-minded the theorist is. In the starkest of interpretations, to shun **RBI** is identified with irrationality and $1 - r$ becomes the probability of behaving irrationally with a view to confusing player B. Alternatively, game theorists may wish to allow $1 - r$ to be the probability with which player A espouses either the logic of

Figure 4.4

or some other logic without requiring that there is a hierarchy of logics with **RBI** at its pinnacle. Indeed, if they are keen to show that the critique in

Section 4.4

is a special case of equilibrium logic, then they must accept that **RBI** is just one of many equally admissible conceptualisations. (Later I will be arguing that this last claim is reminiscent of the postmodern challenge to Reason.)

In the second leg of the defence we find the plausible argument that there are three reasons why player A may choose DOWN at $t = 1$ against the advice of **RBI**.

22

Firstly, she may play DOWN because she is irrational. Secondly, although rational, she may not subscribe to the logic presented by **RBI** and

Figure 4.2

. Thirdly, she may be rational *and* subscribe to the **RBI** (and the logic of

Figure 4.2

) but, nevertheless, attempt to confuse player B through her choice at $t = 1$ so that B plays DOWN at $t = 2$ giving her a chance to reap the highest payoff at $t = 3$. Of course, observation can never help B distinguish between the second and the third reasons. If A is irrational then, potentially, this will be revealed at $t = 3$ where she will choose DOWN. On the other hand, if she is rational and plays DOWN at $t = 1$, then her particular version of rationality will never be revealed via her choices as she will invariably play UP at $t = 3$. Therefore, player B lumps the second and third reasons for a rational A playing DOWN at $t = 1$ under one category and attaches to this event probability $1 - r$. In any case, this is not a serious conceptual problem since one can argue that to doubt **RBI** is conceptually identical to not doubting it and yet rationally to choose to evade it.

Assumptions:

- (a) AbB is rational with probability $1 - w \leq 1$.
- (b) BbA is rational with probability $1 - p \leq 1$.
- (c) A and B know (a) and (b).
- (d) If at $t=2$ p were to equal 1 then: $BbA \rightarrow \text{DOWN}$ at $t=3$ with certainty.

Definitions:

- (e) Let $p' > 0$ be the probability belief of B at time $t=2$ that would induce B to play DOWN at $t=2$.
- (f) Let q be A's probability belief that p exceeds p' ; i.e. $q = \Pr(p > p') \mid \text{she plays DOWN at } t=2$.
- (g) Let $q' > 0$ be the probability belief of A at time $t=1$ that would induce A to play DOWN at $t=1$.
- (h) Let $1 - r$ be the probability belief of B that A, if rational, will adopt the RBI reasoning (i.e. the reasoning of Figure 4.2).
- (i) Let s be the probability belief of B that $A \rightarrow \text{DOWN}$ at $t=1$ when A is irrational (or differently rational).

Fundamental conjectures:

- (j) A rational player A believes that if she defies the logic of backward induction (RBI) and plays DOWN at $t=1$, then B will revise p upwards at $t=2$ using Bayes' rule – see equation (4.1).
- (k) $q > q'$ at $t=1$ and, therefore, $A \rightarrow \text{DOWN}$ at $t=1$.

where b and \rightarrow denote the verbs *believes* and *plays* respectively.

Figure 4.5

The positive defence: a game theoretical explanation of the deviant strategy.

Let us now explore the interdependence between agents' beliefs and choices.

Figure 4.6

captures the possible states for player A at $t = 1$ as perceived by player B. Player B expects A to be rational with probability $1 - p$, in which case she may adopt **RBI** with probability r or choose an alternative logic with probability $1 - r$, or to be irrational with probability p . Probability s relates the likelihood that an irrational A will choose the deviant strategy at $t = 1$.

Suppose now that player A plays DOWN at $t = 1$. What should B think? As in

Figure 4.4

, he will immediately update his probabilistic expectation that A is irrational taking into account the possibility that she may simply be using an alternative reasoning to **RBI** (such as the one in

Figure 4.4

). Bayes' rule recommends the following consistent updating mechanism once A's behaviour is observed at $t = 1$.

$$\begin{aligned} \Pr(A \text{ is irrational} \mid A \rightarrow \text{DOWN at } t = 1) \{ = p_2 \} \\ = \frac{\Pr(A \rightarrow \text{DOWN at } t = 1 \cap A \text{ is irrational})}{\Pr(A \rightarrow \text{DOWN at } t = 1)} \end{aligned}$$

[where the subscripts of p correspond to the time period at which these beliefs are formed]

From

Figure 4.6

, it follows that the above updating mechanism can be re-written as

$$p_2 = \frac{p_1^s}{(1 - p_1)(1 - r) + p_1^s}$$

Hence, given values for r and s , both players can work out how an initial belief that A

is irrational will be updated if A plays DOWN at $t = 1$. Moreover, they both know the value of p' (i.e. the degree of conviction at $t = 2$ that A is irrational so that B wishes to play DOWN at $t = 2$) and they can work out whether it would make sense for a rational A to play DOWN in order to ensure that p reaches p' (i.e. so that deviant behaviour at $t = 1$ pushes p_2 up to the level of p'). In our particular game in

Figure 4.1

, p' equals $1/11$. Given an initial probabilistic belief by A that B is irrational equal to p_1 , the above Bayesian updating formula is re-written as:

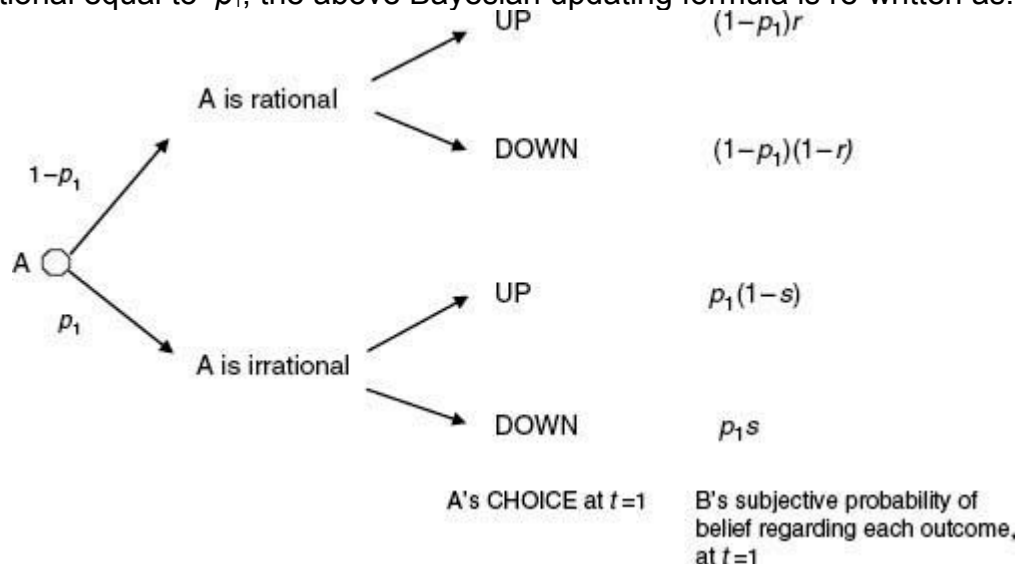


Figure 4.6

Player B's conjectures about player A at $t = 1$.

$$\frac{1-p_1}{p_1} = \frac{1}{\gamma} \frac{1-p_2}{p_2}$$

where $\gamma = (1-r)/s$

23

Our players know that if 'A → DOWN at $t = 1$ ' is to create a significant degree of doubt in B's mind concerning A's rationality, p_2 must reach at least $1/11$. Supposing that, for example, $1-r = 1/2$ and $s = 1/4$, what is the minimum probability belief at $t = 1$ with which B expects A to be irrational? Substitution into (2) yields this level as $p_1 = 1/21$. In summary, the above model tells the following story:

If $p_1 > 1/21$ then A knows that B will be prepared to risk playing DOWN at $t = 2$ if she plays DOWN at $t = 1$

24

If $p_1 < 1/21$ then A knows that B cannot be made to feel with sufficient strength that A is irrational. So, unless A is irrational she will play UP at $t = 1$

If $p_1 = 1/21$ then A is indifferent between the two strategies at $t = 1$ and randomises. If the outcome of this randomisation is DOWN then p_2 will (by equation (2)) equal $1/11$ and B will also become indifferent between his options at $t = 2$. Thus, he will also randomise.

We have come to the end of an impressive defence of game theory built on standard game theoretical concepts developed by, amongst others, David Kreps and Robert Wilson, Paul Milgrom and John Roberts (see Kreps *et al.* 1982). It claims that once the **CKR** assumption is dropped (i.e. once $p_1 > 0$), player A may attempt creatively to exploit the fact that her opponent does not know if she is rational at all or, if rational, what kind of rationality she espouses. In addition, she may have an incentive to defy **RBI** even though she initially conceived of the game in terms of **RBI**! Turning the tables

on the challenge of

[Section 4.4](#)

, game theory seeks to explain internally the logic of

[Figure 4.4](#)

by means of

[Figure 4.5](#)

The only striking difference between this sophisticated narrative and that of

[Figure 4.4](#)

is that the latter begins with an assumption in common and absolute belief in each other's rationality, whereas the former requires that B experiences at least a little bit of uncertainty concerning A's thoughts. One could construct a claim that

[Figure 4.4](#)

is superior to the above as an explanation of deviant play at $t = 1$ because it allows deviant thoughts even when both players are convinced that they share the same reasoning. However, on its own, this would be a thin claim. For the defenders of game theory could retort that, in view of the conclusions of

[Figure 4.4](#)

, no rational player can be certain that she knows the reasoning that her opponent will employ. Thus, assumptions (a) and (b) in

[Figure 4.4](#)

are not the assumptions that rational players would wish to make and, consequently, it makes more sense to relax the **CKR** assumption instead, as in

[Figure 4.5](#)

4.5.1 A repudiation of the positive defence of equilibrium theory

At the centre of the positive defence, above, we find Bayes' rule. It provides the link which was missing from

[Figure 4.4](#)

and allows the conclusions of the logic in that table to hold without compromising the logic of equilibrium. Its role is to update B's initial concern about A's possible irrationality after A plays DOWN at $t = 1$. However, I wish to argue that its use in this context is so fraught with problems that we (and our players) are better off without it.

²⁵

If this turns out to be sound advice, we will return to

[Figure 4.4](#)

and the equilibrium defence will have been fruitless as there will be no unique (equilibrium) story to tell about how our players process the information that deviance at $t = 1$ furnishes.

What conditions must hold for Bayes' rule to be operational? Assuming that A and B share the same rationality (i.e. **RBI**), player B must know the values of p_1 , r and s . Then, B must believe that player A knows that he knows these probabilities, which requires that A and B have exactly the same subjective probabilities on p_1 , r and s . If such convergence of minds is not achieved, player B will not be able to use the Bayesian updating mechanism above, and A will not expect him to. Let us take these subjective probabilities one at a time:

1A and B have the same expectations: i.e. they have somehow homed in on the same value of the probability that an irrational person will play DOWN at $t = 1$. But is this a reasonable deduction from the assumption that A and B are both rational? Surely the point about irrational or stupid agents is that rational agents *do not* understand them. Even if one is convinced one can predict irrational behaviour (i.e. form a sound estimate of s), how can one be sure that another rational agent will form *exactly* the same

estimate?

26

And what happened to the newly found open-mindedness which would allow for more than one kind of rationalisation? If this concession is genuine, then surely there must be more than one commentary on irrationality thus giving rise to a plethora of predictions on s thus wrecking Bayes' rule.

2A and B share the same value of r : i.e. if A and B are both rational (regardless of the particular form of rationality they subscribe to), B knows the probability with which A will play DOWN at $t = 1$. In effect, the positive defence postulates the existence of a unique theory by which player B can predict or explain the behaviour of someone whose reasoning he does not share. But how is this possible when there are many possible modes of reasoning? Moreover, even when they share the same r how can A be absolutely certain that this is so? For that is exactly what is required before Bayes' rule can function.

3A and B share the same value of p_1 : i.e. a rational player A must know exactly B's subjective probability assessment that A is irrational and must know that B knows that! That they are rational when asked to form this identical belief is no guarantee that they will form it. Once more the assumption of rationality is asked to do too much.

CRITICISM 4.1 *The positive equilibrium defence refurbishes the common knowledge of rationality assumption (CKR) only this time it is common knowledge of p_1 , r and s – not of rationality as in*

Figure 4.2

. Since no one can demonstrate that equally rational agents ought to trust each other to have the same subjective beliefs p_1 , r and s , it would be utterly irrational of them to act in a way that vindicates the theory proposed by the positive defence of equilibrium analysis. Therefore, I conclude that the challenge posed by

Figure 4.4

has not been met by game theory.

The standard reply by game theorists which the above criticism shall occasion is that if we want agents to entertain different expectations, then there is no real problem. We assume that this is the case, equip them with probability density functions which capture their uncertainty about each other's subjective beliefs and we derive a complex asymmetric information model that addresses the above concerns. However, this would muddy the waters unnecessarily. There is no gain to be had if a problem is elevated to a higher level of complexity without being solved. For if an equilibrium model with asymmetric subjective beliefs is to work, we must assume that the probability density functions of one player are known *with certainty* by another. So, instead of demanding that agents use the same value of p_1 , r and s , the positive defence now demands that they are certain that they have the same probability density functions over different values of p_1 , r and s . But why would they feel confident that this is so? Such confidence would be unacceptable on the grounds of rationality.

27

In order to avoid the danger of getting bogged down in a pointless argument about higher order probabilistic conjectures, and whether or not they should be in equilibrium, let me make a concession that I do not have to make and yet demonstrate that the common knowledge the positive defence depends upon is implausible. Suppose that A and B are equally rational and have at their disposal exactly the same values of p_1 , r and s , as the positive defence assumes (i.e. for the moment disregard *Criticism 1*). Some unspecified process leads them both to the conclusion that $1 - r = 1/4$, $s = 1/2$ and $p_1 = 1/21$. The question then is: what will A and B do given these shared beliefs? Will they have an incentive to move away from them? According to the equilibrium story, if player A is rational she must randomise at $t = 1$.

28

If the outcome of this randomisation is DOWN then Bayes' rule yields $p_2 = 1/11$ and player B is forced to randomise at $t = 2$ too. This is a knife-edge situation where neither has an equilibrium pure strategy and where each has to resort to an equilibrium mixed strategy – i.e. to randomising. Is this what they will do?

Suppose player A believes that B will stick to the above scenario. If she does, she has no overwhelming reason for playing DOWN, UP or for randomising. So, why should she randomise? Her expected returns are the same whatever she does and, hence, she may choose one of the two strategies with certainty or indeed choose to mix them in any which way she feels like. Suppose that for some unspecified reason she contemplates playing DOWN.

DEVIANT THOUGHT 1 (DT1). A decides to set $r = 0$.

Naturally, DT1 is not an equilibrium decision since only $r = 3/4$ would ensure that her choice will be in equilibrium with B's conjectures. On the other hand, it is not a foolish decision either since whatever r is set equal to, her expected returns are the same provided B believes r to equal $3/4$. To put it differently, A has no incentive to stick to $r = 3/4$ even if this is the value she initially entertained. Thus, if she espouses DT1, she will choose DOWN at $t = 1$. Now, what if B thinks that there is a tiny probability that A has adopted DT1? He will immediately set $r = 0$ in the Bayes rule formula above and derive a new probability estimate concerning A's irrationality (i.e. a value for p_2) that is below $1/11$ and is incapable of motivating him to play DOWN at $t = 2$. This is captured by the second deviant thought:

DEVIANT THOUGHT 2 (DT2). B anticipates DT1 and if $A \rightarrow \text{DOWN}$ at $t = 1$, $B \rightarrow \text{UP}$ at $t = 2$.

Not surprisingly, a string of deviant thoughts may follow. Player A may anticipate that if she plays DOWN then DT2 will emerge in the mind of B and she may, therefore, set $r = 1$ since she expects B to play DOWN at $t = 2$.

DEVIANT THOUGHT 3 (DT3). A anticipates DT2 and sets $r = 1$.

If player B expects DT3 to infiltrate A's thoughts, then we move to DT4:

DEVIANT THOUGHT 4 (DT4). B anticipates DT3 and sets $p_2 = 1$

if A plays DOWN at $t = 1$. Hence, B will be prepared to play DOWN at $t = 2$. If player A thinks that playing DOWN at $t = 1$ will give rise to DT4, then she will develop DT5:

DEVIANT THOUGHT 5 (DT5). A anticipates DT4 and sets $r = 0$. And so on.

CRITICISM 4.2 *It is not only that players will converge on the same subjective probabilities by accident alone but, moreover, that they may busily develop thoughts which will ensure the impossibility of such symmetry.*

Since the actual outcome of the game will depend on which thought each player terminates his or her climb up the ladder of conjectures, we cannot predict what either of them will do. There is no optimal stopping rule when one enters deviant trains of thought and, for this reason, any equilibrium story (including those postulating probability expectations over the two strategies of each player) is inappropriate.

29

All we can safely say is that, if A stops at DT3 then

[Figure 4.2](#)

applies. If, on the other hand, she reaches DT5, then

[Figure 4.4](#)

aptly describes her thoughts.

I predict two neoclassical objections to *Criticism 2*. First, DT1 may be denied on the grounds that there is no reason why it is more likely to develop than, say, DT1': A sets $r = 1$. However, in this case player B may anticipate DT1' and move

directly to DT4 setting $p_2 = 1$. The point is that at $t = 1$ anything goes whichever deviant thought arises first. The second objection is that *Criticism 2* applies only when p_1 , r and s are such that $p_2 = 1/11$. Although this is correct, this case is too important to dismiss as an exception that confirms a rule. Since A's choice at $t = 1$ will depend

on whether she expects or not $p_2 > 1/11$, it is crucial for the positive defence that there exists one combination of p_1 , r and s such that $p_2 = 1/11$ so that A is made indifferent between UP and DOWN at $t = 1$. Otherwise there is no clear demarcation between the case where A will rationally play the deviant strategy and the case where she will not. And without this demarcation, there can be no equilibrium defence of game theory from the challenge of 4.2. In conclusion, *Criticism 2* reinforces the claim of *Criticism 1* that there can be no tenable equilibrium theory of what will happen if A and B are gifted with equal amounts of Reason.

30

At this stage it is helpful to summarise the argument in simple terms.

Figure 4.4

introduced the possibility that agents will contemplate a risky strategy. The positive defence tried to explain this as an equilibrium strategy. However, the moment such a strategy is contemplated, there is no equilibrium solution. When the

Figure 4.4

strategy is considered, an agent's choice depends on subjective judgments about another agent who must himself make subjective judgments about her earlier and future behaviour. The nature of agents' belief formation being subjective, it undermines the derivation of equilibrium solutions. Quite clearly, equilibrium theory survives only if somewhere along the line we *assume* an equilibrium outcome. Its positive defence, if it is to remain erect, needs to be underpinned with the hidden statement: 'Let us assume that equilibrium theory is correct.' But this would be equivalent to the negative defence that its positive variant was meant to improve upon!

4.6 Postmodern, Humean and dialectical interventions

An eagerness to unravel logically complex social phenomena is a commendable characteristic of modernity. The problem is that, along the way, its ambition often sweeps unresolved questions under the carpet in search of short cuts. Postmodern thinkers have questioned the concepts that modernity uses on the grounds that they are flimsy analogies rather than concepts. One of the concepts that they challenge is that of Reason. Those who are concerned about game theory's rationality postulates may find the postmodern critique useful. Looking at the preceding arguments through a postmodern prism, one interpretation of the discussion in the preceding sections is that **CKR**, the assumption of a common knowledge of rationality of infinite order, is an extreme form of modernity; a by-product of an illegitimate, yet strong, ambition to select one of A's two strategies at $t = 1$ as uniquely playable by rational agents. Postmodernists would recognise in **CKR** (and

Figures 4.3

and

4.4

) the same modernist tendencies that they disparage in literary criticism, politics and philosophy (see Derrida 1978; Norris 1985; Lyotard 1984).

It is, however, perfectly admissible to accept the critique of game theory without abandoning modernity. In a Humean sense, Reason is the slave of passions (that

is, the payoffs in our game) which motivate choices and acts as the disinterested judge who weighs the merits of the various options but does not pass judgment on the desires themselves (in the same way that a judge does not question the law). If desires under-determine choice, Reason is not to blame for the resulting indeterminacy. As Aristotle put it in *Nicomachean Ethics*, the rules of the undetermined are themselves undetermined. Thus, the critique of game theory does not challenge modernity as such but only an extreme version of it (i.e. equilibrium game theory) which wants a determinate solution so badly that it contrives rationality concepts (**CKR**, **RBI** etc.) that are not supported by Reason. Humean instrumental Reason offers no guidance to A

and B at $t = 1, 2$ because no choice is uniquely rational. What it can do is to suggest that, in such circumstances, the solution lies in convention. Conventions help agents make sense of logically indeterminate situations although no convention in itself can be understood in terms of its rationality. If we wish to understand how they are formed, we need to look at their evolutionary stability. Further, if we wish to explain why they become stable, a Humean interpretation is possible: agents develop the desire to follow the established conventions. Then, it may be rational to act in one way rather than in another as a new desire has been actuated allowing Reason to discern a uniquely rational action.

Granted that modernity is not directly challenged by

[Sections 4.3](#)

and

[4.5](#)

, it is worthwhile to follow the postmodern critique of it a little further. Hume shares with game theory a perception of Reason as a concept which is definable axiomatically and independently of social interaction. Postmodernity on the other hand denies the possibility that abstract signifiers such as Truth, Being and Reason signify anything concrete; that they are more than figments of our language. By contrast, we are encouraged to recognise that Reason appears as a momentary flickering of presence and absence and does not allow us a good look. The only strategy that we should contemplate is to deconstruct narratives such as the one in

[Figure 4.2](#)

, to invoke Reason, and then immediately to erase and fragment it. In the context of the earlier discussion, we are asked to accept the rationality of

[Figures 4.2](#)

and

[4.4](#)

simultaneously, not because (as the Humean would argue) Reason cannot deliberate in this case, but because there is no such *thing* as Reason.

For a brief moment, the positive defence of game theory in

[Section 4.5](#)

seemed compatible with postmodernity. As it begins with a recognition that there is no unique reasoning and thus no hierarchy of logical trains of thought, one may think that postmodernity has found a mathematical expression. However, the deconstruction of that defence (see the two main criticisms in that section) reveals the inherent incompatibility between the two. For if postmodernity accepts this model, it will be using the concepts it is critical of in order to castigate them and would become vulnerable to a critique reminiscent of Heidegger's attack on Nietzsche.

31

On a positive note, postmodernity offers an interesting answer to a question we have neglected so far. Returning to the first stage of our game, why should we discuss the rationality of various types of reasoning in terms of the backward induction logic? Why should, in view of our conclusions, label

[Figure 4.4](#)

'deviant' thus crediting

[Figure 4.2](#)

with a priority it should not have on the basis of Reason? Postmodernity has this to offer: As the Enlightenment sought scientific

explanations by which to escape dogmatic certainties, natural science took it upon itself to furnish them. In the realm of natural science, the various possibilities that required analysis were states of nature and could be treated as such quite legitimately. When social phenomena were tackled, it was natural to try to apply the same logic. The problem is that human choices cannot (and should not) be treated as states of nature.

Take for instance backward induction. If A was to play the game not against a human agent but against an automaton whose software was describable by

Figure 4.3

, then backward induction would correctly inform her that she should play UP at $t = 1$. However, when she plays against a human B, backward induction breaks down. Nevertheless, the cognitive priority that we seem to lend the backward induction logic is, according to postmodernity, a historical accident. It is simply the product of what it mockingly refers to as the 'Enlightenment episode'.

What picture of the agent is postmodernity drawing? It looks at our game and observes that at $t = 1, 2$ modernity offers no useful commentary. Only when the game reaches (if it does) $t = 3$ does modernity have an answer: A will play UP. But what kind of human subjectivity does this imply? Human creativity is responsible for creating and simultaneously undoing the backward induction logic

and is capable of frustrating all attempts to treat agents as automata. If we want a metaphor for understanding postmodernity's view of subjectivity, imagine the individual as a multifaceted and disintegrating interplay between selves; a series of different masks. Instrumental rationality is an empty concept if one espouses this model of men and women.

Lest we wrongly conclude that postmodernity be the only alternative to the Humean perspective, it is valuable to look at the contribution of Hegel. At the risk of oversimplification, a Hegelian interpretation of what is happening at $t = 1$ in our game is best portrayed in juxtaposition to the Humean and the postmodern views. The former evokes the image of a static Reason which, due to the inability of desires to provide it with enough information, stays on the sidelines and refuses to engage until $t = 3$ is reached, whereas the latter agrees that Reason is absent at $t = 1$ but claims that this is due to its non-existence. By contrast, Hegel would argue that Reason jumps into the fray at $t = 1$ and generates *contradictory* thought processes like those in

Figures 4.2

and

4.4

. And there is the rub. For it is these inconsistencies that give Reason the opportunity to enrich itself with elements of fundamentally opposed reasonings. Hegel views Reason as an evolving concept that affects the agents' experience and, in contradistinction to Hume, is affected by it.

Looking at our little game again, Hegel's dialectics suggest that at $t = 1$ Reason generates two contradictory logics (

Figures 4.2

and

4.4

) which are equally powerful; they are the thesis and the antithesis. The outcome is only describable in historical (as opposed to logical) time because of the logical equivalence of the two types of reasoning. However, once the game is played (and here Hegel would agree with Humeans and postmodernists that there is no way of predicting what will happen if all the information we have is in

Figure 4.1

), the Reason of agents, as well as of theorists, emerges superior to what it was before they encountered this

game. As it absorbs both logics (

Figures 4.2

and

4.4

), Reason endows us with an understanding of the game that is indescribable by one of the two figures although it is comprehensible by a synthesis of the two.

Put differently, our rationality was of a lower order of development before we stumbled on this game. Generally, the more complex the social phenomena to which men and women are exposed the more advanced their Reason. Reason develops as rational agents struggle to come to grips with the maze of conjectures that social interaction (of which our game is a simple example) creates. To use Hegelian language, rationality is not to be defined axiomatically but is to be understood as a *process*.

34

If we wish to follow modernity in picturing Reason as a totality, we may still do so. However, it is not a static totality but one whose aspects are in contradiction with each other. And through this internal feud, the aspects of the totality (e.g. the conjectures in

Figure 4.2

or

4.4

) transform not only the totality but also each other.

35

It is quite obvious that Hegel and Hume are on modernity's side. Excepting their disparate language, what is the significant difference between the two interpretations? Both would accept the indeterminacy at $t = 1$ of our game and neither would deny

Figures 4.2

or

4.4

their respective worth as equally plausible conceptualisations. I suggest that the main difference lies in what we may describe as the by-product of the indeterminacy. Following Sugden's (1989b, 1991) reading of Hume, the by-product is the convention that will help agents choose in the absence of abstract logical guarantees. Reason does not shape these conventions itself, although they are compatible with it. What gives rise to an impetus for their generation is the need to serve existing desires. Furthermore, in order to entrench the fledgling conventions, a new desire to abide by the evolutionary stable convention evolves – the birth of morality (or normative beliefs). It is this new desire that unlocks the problem and breaks the indeterminacy. However, the driving force behind such evolution is (a) given desires and (b) an unchanging Reason. More importantly, in this model it is impossible to pass judgment on the rationality of social conventions since Reason has had nothing to do with the selection of the particular convention. It is also futile to imagine that there is some overarching social goal that guides the evolution of conventions.

In contrast, in the Hegelian perspective indeterminacy bears, in addition to new desires, a new mode of reasoning – a fresh conceptualisation of one's self as one encounters the 'other' in a social setting. Whereas in Hume indeterminacy actuates conventions and possibly new desires, in Hegel it also actuates an upgraded version of Reason. Desires, beliefs and Reason change at once when agents meet each other in a society that brings them face to face with profound contradictions. Desires and Reason are thus endogenously produced social products. The major implication is that Hegel, as opposed to Hume, sanctions judgements of the rationality of social norms that 'solve' social games on the basis of a historical analysis. When we look at past conventions we are at liberty to castigate them even if it is possible to show that agents who abided by these conventions did so because their Reason could not determine otherwise what they ought to do. Since Reason progresses in historical time, conventions that were spawned

by a previous set of social circumstances, and which were perhaps compatible with agents' rationality *at that time*, may not pass the test of Reason today. While Hume's philosophy does not allow us to pass moral or political judgment on social conventions,

Hegel's does.

4.7 Epilogue

This chapter focused on an ultra simple game involving two rational agents. What is quite astonishing is how much can be gleaned from such a simple interaction about the state of economics.

We saw how economists, using game theory (neoclassical economics' highest form), struggled to show that their analysis 'solved' the game; that they could produce the definitive narrative on how rational players would play this game. To do so, they demonstrated that the game possesses only one 'equilibrium'; i.e. only one strategy per player such that player A's strategy is, inter-temporally (i.e. at each one of the game's three potential stages), the best reply to B's strategy and vice versa. The problem was (as this chapter's challenge to their 'solution' showed) that rationality cannot compel rational players to stick to that equilibrium game-plan. Indeed, the one solid conclusion of the previous sections was that, even in this simple game, there can be no unique prescription about how the game ought to be played by rational agents. In short, indeterminacy rules.

Yet, economists resist this simple truth doggedly and with impressive determination.

Section 4.5

is rather instructive of the panache, sophistication and logical machinations with which economists struggle to defeat indeterminacy; to insist that their equilibrium analysis, if allowed to scale magnificent heights of complexity, can, in the end, prevail and offer a complex, stochastic but nonetheless definitive account of the precise strategy that rational agents *must* employ. Of course, in order to 'prove' this point, in order to convince the world that they have bested indeterminacy, and that they have procured a uniquely rational strategy per player, they are compelled to usher in, through the backdoor, hidden assumptions that defy rationality, and even common sense.

The question is: Why struggle so tenaciously to solve an unsolvable problem? To pretend that they have rendered determinate an indeterminate interaction? To counter Aristotle's subtle point that the rules of the undetermined must themselves be undetermined? The simple answer is that, just like the good salesperson succeeds through 'closing' deals, so too neoclassical economics draws its discursive power from 'closing' models. Whatever it takes! Even if it means bending the rules of logic and introducing inferences that common sense rejects. When analysing models of entry deterrence in oligopolistic settings, of Central Bank credibility (e.g. against meddling politicians or recalcitrant private sector banks), of negotiations in the context of the World Trade Organisation, etc., neoclassical economists must somehow show that their analysis sheds bright light, uniquely, on the outcome of these interactions. To show this, they follow a two-step method.

First, they work out the parties' equilibrium behaviour. Secondly they *imply* that, to the extent that this equilibrium path of behaviours is well-defined and unique, rationality *compels* agents to stick to this equilibrium behaviour. As long as the economists' audience accepts this implication (that rationality compels agents to stick to the equilibrium path), economists can bask in the glory of having produced the best narrative possible on the outcome of the phenomenon under study (e.g. market structure, Central Bank strategy, WTO agreements). The trouble, of course, is that a sophisticated reader of their work has no reason whatsoever to accept their implication that rationality necessarily compels the rational to stick to an equilibrium path. The analysis of the game in

Figure 4.1

offers an excellent case in point.

The gist of the problem, as unveiled by the

Figure 4.1

game, is that in most strategic interactions, human reason can creatively exploit its own successes and subvert its own logic, thus yielding an indeterminate outcome. In the game we spent this chapter analysing, a first stab at deciphering its strategic structure leads to the conclusion that player A, if rational and respectful of player B's rationality, will play DOWN at the outset. But if A knows that B knows this, A has good cause to consider what B might think if A does the opposite of what rationality demands of her. For if A surmises that B will be 'thrown' by this 'deviation', she may decide rationally to behave seemingly irrationally. Or, put plainly, to bluff. Like all bluffs, it may work or it may not work. Regardless of whether player A bluffs or not, or whether her bluff succeeds, the very possibility of a rational bluff makes it possible for rational behaviour to cross into disequilibrium territory; in which case, the neoclassical analysis that delineates the equilibrium path is rendered irrelevant.

Bluffs are a potent tool in the hand of human agents who may benefit from instilling doubt in the mind of their opponents not only of how strong their hand is but, importantly, of how rational, calculating, emotional or downright deranged they may be. Bluffs make it reasonable to consider pretending to be irrational. Any attempt to model this process by means of a set of equilibrium strategies is bound to prove indeterminate unless the theorist is so determined to define the rational outcome that she arbitrarily bans the very possibility of a successful bluff. This is, indeed, what neoclassical economists do. Of course, they never admit to it. Instead, they introduce a concealed assumption that all probabilistic beliefs must be consistently aligned, even while bluffing; which, of course, is the same thing as to assume that bluffs cannot work.

It is, indeed, impressive that the self-proclaimed purveyors of rational analysis resort to such cheap tricks in order to convince their audience that they have pinned down the unique outcome of an indeterminate interaction.

VERDICT: Economists built a whole array of powerful theories on the game theoretical logic of backward induction and equilibrium behaviour which allowed them to 'solve' complex strategic interactions in all sorts of fields, from industrial organisation, to monetary theory, bargaining models (see also the previous chapter), international trade etc. The problem is that the very kernel of this logic suddenly became the subject of a decisive logical challenge (see

[Section 4.4](#)

). The essence of that challenge was the paradoxical thought that rational players may benefit from defying the equilibrium strategy that, supposedly, maximises their benefits.

This was, in terms of

[Chapter 1](#)

's *dance of the meta-axioms* diagram in

[Figure 1.1](#)

, the challenge which threatened the profession's capacity to 'solve' all these models. Interestingly, it was a challenge that game theorists valiantly took on, and tried to address (arrow a in the same diagram). Alas, despite their best and most impressive efforts, it was impossible to avoid the usual crash on the Wall of Indeterminacy. As

[Section 4.5](#)

showed, desperate analytical attempts to incorporate the theoretical challenge within the neoclassical equilibrium narrative could only succeed if the theorist deployed the strict version of the third meta-axiom (methodological equilibration). In short, by means of a veiled axiom, neoclassicists imposed, against all norms of rational deduction, a strict alignment of beliefs that rules out rational bluffs and, therefore, renders the analysis far more complex but just as unrealistic as it was before the theory attempted to meet the challenge. Summarising, in the parlance of the *dance of the meta-axioms* diagram, the profession (just as in the case of

[Chapter 3](#)

) followed path **a**, but finally reverted to backslide **b** toward a version of its original position that is even less realistic and certainly more indefensible than the original equilibrium theory. Thus, we observe another case of the **1→2→3→4shuffle**.

Lastly, as with every chapter epilogue in this book, the question arises: How has this 'shuffle' enhanced neoclassicism's discursive power in spite of its radical theoretical failure? The answer is obvious: reread

Section 4.5

. It takes a long immersion in its type of logic in order to discern the neoclassicists' sleight of hand. 'Outsiders' (i.e. reasonable people who have invested their energies in other intellectual enterprises) stand no chance of working out the neoclassicists' unwholesome 'trick'. Only graduate students and young academics who have wasted their youth trying to become neoclassically trained economists may ever grasp how these models hang together (and how logically incoherent they are). Once they catch a glimpse of this, they face a stark choice: Write papers in which they do not employ the neoclassicists' sleight of hand, in which case their models will not be 'closed' and their papers will remain unpublished; or write papers which silently employ the same inappropriate axiom, get them published in decent journals and enjoy the rewards of having been inducted into the priesthood. It is clear that a Darwinian process is in play which reinforces the neoclassical method within the economics profession, keeps the hidden axioms out of sight, and produces weighty undergraduate textbooks full of economic models training young economists to 'solve' models without ever understanding the logical incoherence of the solutions that they derive skillfully and, indeed, proudly.

Notes

- 1 Since they can no longer argue that the failure of their theory to predict strategic behaviour is necessarily the result of people's substandard rationality.
- 2 Here I am referring to a specific model of bargaining that has come to dominate the literature: the so-called Nash bargaining solution – see [section 3.4.2](#). Also recall that in game theory, as in most economic analyses, to be rational is to know how to deploy your means effectively in order to achieve your ends. Rationality is exclusively instrumental.
- 3 See, for instance, Rubinstein (1982). In his model the distribution will deviate from the 50–50 division to the extent that one player issues her demand *before* the other. As the delay between demands vanishes, the equilibrium outcome tends to be a 50–50 split.
- 4 One can, perhaps, accommodate the postmodern view in terms of the computer parallel. Before the theory of chaos, one expected the same algorithm to give identical results if fed the same initial values twice. Since the study of non-linear models has revealed that, because the input *can never be exactly the same* twice, the output may be drastically different. So, why should we expect our two agents to come to the same conclusion? If they espouse minutely different conventions by which to predict the thoughts of others, their train of beliefs may lead them to seriously different conclusions and, thus, disagreement.
- 5 See Finelli (1990) who traces the debate on the nature and role of irreconcilable oppositions to the Sophists.
- 6 Recall that game theory does exactly this. It starts with simple games, such as the simple bargaining problem in [Chapter 3](#) or the game of [Figure 4.1](#) here, and 'derives' solutions for them. Once this stage is over, it then looks at more complex situations (e.g. asymmetric information) and uses the earlier assumptions to obtain explanation. This is what I call the analytic-synthetic method of game theory.
- 7 This is a variant of a game that appears quite often in discussions of game theory. See, for instance, Binmore (1987) and Sugden (1989, 1991). For the purpose of easier exposition, I assume that A is female and B is male.
- 8 To be precise this is the so-called subgame-perfect Nash equilibrium. A Nash equilibrium is an outcome brought about by strategies which are chosen on the basis of beliefs which are *ex post* confirmed by the outcome. The equilibrium

is subgame-perfect if the game comprises more than one stage and such a coordination of strategies and beliefs (i.e. an equilibrium) is achieved not only for the whole game, but, also, in each subgame.

9

Binmore (1987) criticises over-reliance on backward induction, Sugden (1989) shows that it is possible to have a game theory without this kind of induction provided we are less ambitious and Pettit and Sugden (1990) cement the arguments against the logic in

Figures 4.2

and

4.3

. More recently, Sugden (1991) provides a good summary of the case against backward induction.

10

Thinking about the possibility of defying the theory that is supposed to govern one's behaviour, is a uniquely human capacity. It is also a capacity that makes the life of the social scientist inordinately demanding. To disallow counterfactuals within a theory (which is what

Figure 4.2

does) is to ask for serious trouble since human rationality has the bad habit of instructing agents to ask, 'what if I do not obey the theory's rules?'. In

chapter 6

of Varoufakis (1991), I argued that counterfactual reasoning is, at once, rational *and* incompatible with equilibrium game theory.

11

The reader may notice that I have made a rather strong assumption concerning what B expects an irrational A to do. Indeed, I assume that an irrational A always does the opposite of what is good for her. This has allowed us to assume that if $p = 1$ – i.e. if B is convinced that A is irrational – then he expects her to play DOWN at $t = 3$ with certainty. This is, of course, too restrictive. Nonetheless, the main point I am making is not lost if the assumption is relaxed. Suppose, for instance, that an irrational A chooses *as if* by randomisation. Then, at $t = 3$ an irrational A plays UP or DOWN with probability 1/2. In this case, p and q can be re-computed fairly easily and the argument remains intact.

12

Those familiar with game theory may protest that game theorists recognise the legitimacy of a logic such as that in

Figure 4.4

without abandoning game theory's tenets – for example see Kreps *et al.* (1982). This is correct. However Kreps *et al.* (1982) can only do this after they assume right at the start that agents have some doubt about the rationality of their opponents.

Figure 4.4

by contrast does not require such a dilution of the common knowledge of rationality assumption: non-equilibrium strategies are rationalised

even when everyone is (at the beginning) absolutely sure that all others are perfectly rational. The ideas in Kreps *et al.* (1982) become relevant in section 4 in which they help construct a defence of game theoretical orthodoxy.

13

The reader may ponder the generality of my conclusion in view of the fact that I have focused on a single game. Is it fair to discuss the whole project of game theory on the basis of one example? I think it is. For this is an example that contains a unique Nash equilibrium (subgame perfect) which should, if game theoretical thinking is to be vindicated, produce an unequivocal rational strategy (due to the uniqueness of the equilibrium). If the logic of

Figure 4.4

is compatible with full rationality, then we have evidence that the existence of a unique equilibrium does not necessarily tell us what agents will do. Since game theory trades on the thought that it ought to, one example where this is untrue is as good as a thousand.

14

The *Harsanyi doctrine* occupies a central role in game theory since on it rest a very large number of solutions that would otherwise break down. Hargreaves-Heap and Varoufakis (2004) discuss extensively the *Harsanyi doctrine*, its implausible nature and game theory's reliance on it. In the present context, the negative defence draws on it heavily.

15

Suppose for instance that $q > 1/50$ and A plays DOWN at $t = 1$ but that B sets p at 1/20 and plays UP at $t = 2$. Alternatively, suppose that $q > 1/50$, A plays DOWN at $t = 1$, B sets p equal to 1/8 thus playing DOWN at $t = 2$ and, finally, A plays UP at $t = 3$. In both these cases one of the two has formed expectations that are proven erroneous *ex post*.

16

This is effectively the thesis in Bernheim (1984).

17

Recall the earlier argument that the equilibrium logic is perfectly legitimate even if not uniquely so. Thus, a player may still choose to be prudent and assume that, since her opponent is equally rational, there is nothing she can do to confuse her.

18

Figure 4.2

, for example, depends on the unique projectibility of traits established at $t = 3$ onto $t = 2$ and $t = 1$.

19

The game of

Figure 4.1

is truly interactive in that what player A does at $t = 1$ depends entirely on what A thinks that B will think if It is in

such a game that the enigma of human reasoning becomes pertinent and wrecks the certainty of backward induction. In other cases, where the choice of one player can be made independently of conjectures concerning the actions of another, then of course induction is straightforward. Consider the following ten dot game. There are ten dots which two players take turns to erase. The first player begins and may erase either one or two dots. Then it is the second player's turn to either erase one or two dots. The player who crosses out the last dot wins. Working backwards, it is clear that the player who plays first has a unique dominant strategy: to erase only one dot at the beginning. In this way, she can be the first to cross out the 4th, 7th and, finally, the 10th dots whatever player B's choices. Backward induction works impeccably in this game because A does not need to consider what B will think if A plays in one way rather than in another. Then, the trait identified at the last stage of the game is uncontentionally projectible to the very first stage when the game commences.

20

Postmodernity actually rejects the very possibility of a concept. For rhetorical purposes, it may argue that analogies are often mistaken for concepts, in order to demonstrate the vacuousness of concepts.

21

The reader who would like to leave open the possibility that an irrational player acts in an unpredictable manner will protest that this is too stringent an assumption. However, the analysis will not change significantly if we envision an irrational player A as someone who chooses between UP and DOWN as if by randomisation. Footnote 10 applies here with equal force.

22

I assume that A believes B to be rational with probability $1 - w = 1$. The model can be easily extended to allow for two-sided uncertainty concerning rationality, i.e. letting $w > 0$.

23

For the updating mechanism to make intuitive sense, $1 - r > s$ – i.e. the probability that A will adopt some logic different to that of

Figure 4.2

(RBI) if rational and thus choose the deviant strategy must exceed the probability that an irrational A will choose the deviant strategy. This is very sensible since otherwise there would be no reason for B to believe that DOWN at $t = 1$ enhances the prospects that A is irrational.

24

Naturally, part (k) of

Figure 4.6

is tantamount to the condition $p_1 > 1/21$.

25

Binmore (1987, 1988) has also voiced concern about the indiscriminating use of Bayes' rule.

26

There is an interesting parallel here with Foucault's (1967) critique of 'modernity's monologue'. Foucault claims that before the triumph of modernity, there used to be a dialogue between rationality and madness. Later, this dialogue broke down and left us with a monologue of rationality on madness. And yet, he goes on, there are dimensions of sense in madness that are missing in what we tend to think of as Reason, or to put it differently, there is a great deal of Reason in madness. Any attempt to evict madness altogether in order to procure pure Reason is, therefore, ill-conceived. The reader who is so inclined may interpret the assumption that there exists a uniquely rational estimate of s as a technical manifestation of illegitimate attempts to cement this monologue.

27

A player's conceptualisation of her opponent's conjectures is, in itself, a theory. To argue that one attaches, via induction, probabilities to different such theories and, in addition, to insist that these probabilities are common property, is philosophically absurd. Peirce (1932) draws the important distinction between the probability of a hypothesis and the probability derived from a hypothesis. He writes:

It may be conceived, and often is conceived, that induction lends a probability to its conclusion. Now that is not the way in which induction leads to the truth. It lends no definite probability to its conclusion. It is nonsense to talk of the probability of a law, as if we could pick universes out of a grab bag and find in what proportion of them the law held good...What induction does...is infinitely more to the purpose.

28

The reason is that if A goes DOWN at $t = 1$, then equation (1) will update B's probabilistic assessment that A is irrational to $p_2 = 1/11$. This posterior belief makes B indifferent between UP and DOWN at $t = 2$. Thus, A anticipates that DOWN at $t = 1$ will make B randomise at $t = 2$, a thought that makes her unsure as to whether she ought to play DOWN at $t = 2$. Consequently, she also randomises at $t = 1$.

29

See Skyrms (1990) for a discussion of deliberational disequilibrium.

30

Figure 4.4

has presented this critique of equilibrium theory implicitly. Let $p_1 = 0$. Then, if player A played DOWN at $t = 1$, equation (1) cannot be defined: an event occurred that B had attached a zero probability to. So, what should B do in such a situation? According to the equilibrium story, there is no answer. Can we speculate that, in the absence of advice by the theory, player B may still revise p upwards (i.e. $p_2 > 0$). If A expects this to happen (and there is no reason why she should not), then she may rationally choose DOWN at $t = 1$. Of course, there can be no equilibrium account of what has happened. Therefore, equilibrium theory is inferior to the account of

Figure 4.4

because rational agents may have an incentive to violate it.

31

Nietzsche wrote: 'What therefore, is truth? A mobile army of metaphors, metonymies, anthropomorphisms; truths are

illusions of which one has forgotten they are illusions...coins which have their obverse effaced and now are no longer of account as coins but merely as metal' *On Truth and Falsity on their Ultramoral Sense* in Levy (1964). However, Heidegger successfully exposed holes in his arguments by demonstrating that Nietzsche needs truth as a concept in order to argue against its meaning. Interestingly, this is also a problem for Heidegger. Finelli (1990) claims that Being is denied by Heidegger and his contemporary postmodernist followers, but that in their philosophy it returns to determine human reality through its loss and emptiness.

32

Sugden (1991) illustrates this point in the context of (i) a critique of Savage's expected utility theory and (ii) the theory of games.

33

Gerhard Adler writes: 'The enigma of creativeness rooted in the irrational, indefinable matrix of man's timeless psyche has held eternal fascination for him and has helped produce the most memorable justification of his status as man' – see the *Foreword* in Kirsch (1966).

34

The social anthropologist Levi-Strauss (1966) defines analytical Reason as the type of logic that develops when humans try to understand natural phenomena of a low order (eg. hydrodynamics as opposed to the concept of time). He thinks that such logic is frustrated when it is called upon to explain social phenomena. The result of this failure is a new kind of Reason which, in Hegelian fashion, he terms dialectical. '... [D]ialectical reason thus covers the perpetual efforts analytical reason must take to reform itself if it aspires to account for language, society and thoughts; and the distinction between the two forms of reason in my view lies on the temporary gap separating analytical reason from the understanding of life. Sartre calls analytical reason reason in repose; I call the same reason dialectical when it is roused by action, tensed by the effort to transcend itself.'

35

Of course, postmodernity is eager to attack Hegel in the same way that it disparaged Hume. The contradiction on which Hegel bases the sublation of Reason is seen as both unresolvable and as unreal. It is unresolvable because Reason is meaningless and, therefore, hardly capable of improving itself. It is unreal because when we talk of *the* contradiction, we fall victims to the inferiority of our language. The latter is forced, through its imprecision, to contrive false categories (such as Reason and Unreason) when, in reality, *the* contradiction is, like truth in Nietzsche, an illusion that we have forgotten that it is an illusion.

5 Bargaining by rules of thumb

When strategic indeterminacy forces the rational negotiator to fall back on myopic rules of thumb

5.1 Prologue

5.1.1 Background briefing

The last three chapters placed under the microscope neoclassical economics' method for explaining a range of economic phenomena at the heart of almost every market transaction: bargaining, conflict and the mutually beneficial agreements that are available to rational antagonistic parties. We found that every attempt to infuse a minimal degree of realism into the neoclassical analysis of these ubiquitous phenomena led to radical indeterminacy; to a set of optimal strategies and outcomes whose size tends to infinity. We also saw how neoclassicism responded to this indeterminacy; how it erected a wall of denial constructed out of the underhanded assumption that all beliefs and actions must, somehow, be in equilibrium. Finally, we discovered that this axiomatic imposition of an equilibrium between beliefs and stratagems is no more than the adoption of the strict version of the third meta-axiom (see

[Chapter 1](#)

).

The present chapter asks a question hitherto un-posed: Granted that the neoclassicists recoil in horror (arrow **b**) when confronted by the indeterminacy following their acceptance (arrow **a**) of the theoretical challenge (arrow **c**) – see the diagram of

[Chapter 1](#)

–, is there an alternative? Could we have taken the analysis further by acknowledging, and embracing, the inevitable indeterminacy (as opposed to fleeing the moment it rears its head)?

[Sections 5.2](#)

and

[5.3](#)

below show that this is entirely possible. That it is perfectly possible to further the analysis, rather than shutting it down and recoiling via the third meta-axiom back to some pristine, yet sterile, neoclassical orthodoxy.

5.1.2 The rest of this chapter

Neoclassical economics purports to model men and women as hyper-rational utility maximisers and to map out the equilibrium path of their actions and beliefs combinations. Its practitioners are quite happy to be accused of assuming too much rationality on the part of agents. Along these lines, they are amenable to the view that, in reality, real people often use rules of thumb, or conventions, instead of working out mentally some optimal behavioural path before setting off on it. In

a sense, they gladly confess to the sin of assuming too much rationality in order to bask in the glory of having managed to work out how humans would behave if they were inconceivably hyper-rational. However, as we saw in

[Chapters 2](#)

to

[4](#)

, these claims on behalf of the neoclassical method need to be taken with a pinch of salt: The behavioural path that they discover is not really a path but a vista of infinite dimensions which effectively licenses all sorts of behavior as consistent with commonly known rationality of infinite degree. In short, anything goes!

But if anything goes, then the analysis offers precisely no guidance on what to do in situations ranging from the simplest (e.g. the game of

Figure 4.1

) to the most complex bargaining settings. Put differently, if whatever one does is consistent with some equilibrium, then one has no clue how to act, or what to expect of a similarly dumbfounded opponent or collaborator. One is, in short, in the dark. Still, one must act, even if clueless on what to do because all her options are part of some rational game plan. At that point, it is inevitable that people grope around, adopt trial-and-error methods for exploring the pros and cons of alternative behaviours, experiment with various patterns of actions, struggle empirically to discern patterns in the behaviours of others etc. In summary, when Reason offers no guidance, due to radical Indeterminacy, rational people resort to rules of thumb.

Granted that this is the predicament of rational men and women when caught up in genuine strategic uncertainty, is there something sensible we can say about the rules of thumb that will spontaneously evolve? Can we offer an analysis of these rules? I think that the answer is affirmative and set out to show this in

Sections 5.2

and

5.3

Section 5.2

is Varoufakis (1996), which investigates the emergence and evolution of bargaining strategies in the context of industrial relations and conflict. Its starting point is the conclusion of

Chapters 2

and

3

, namely the realisation that neoclassical models of bargaining, conflict and settlements are radically indeterminate and that the neoclassical attempts to kill this indeterminacy off (by means of the third meta-axiom) are illegitimate. In four subsections and one appendix, it illustrates how indeterminacy can be embraced and intertwined with an open-ended (as opposed to 'closed') analysis of the strategies rational bargainers follow in the face of such indeterminacy.

Section 5.3

takes its cue from the end of

Chapter 4

. It too focuses on a finite centipede-like game (just like the game in

Figure 4.1

on which the whole of

Chapter 4

was founded) and shows what kind of strategies rational players would adopt if they accepted

Chapter 4

's conclusion that this game is deeply and irretrievably indeterminate. Moreover, in one of its subsections the analysis is generalised to a situation involving more than just two players.

As always, the chapter's final section summarises and links the conclusions to the book's broader theme.

5.2 Bargaining strategies: toward an evolutionary approach

5.2.1 *Conventional neoclassical versus evolutionary approaches to bargaining*

Conventional models of industrial conflict start, as we saw in

Chapters 2

and

3

, with the assumption that the bargainers' uniquely rational beliefs can be worked out in advance. The neoclassical theorist then struggles (unsuccessfully I claimed) to explain instance of conflict, e.g. strikes, as either the result of institutional constraints or of the possibility of irrationality. By contrast, the evolutionary approach adopted here begins with a recognition that bargaining is naturally indeterminate and that, in the absence of a unique model of rational bargaining, conflict-free agreements between rational trades unions and firms reflect the evolution of one out of many possible conventions.

This subsection explores the alternative interpretation of strikes afforded by this perspective. In particular, it shows how strikes help shape the dispositions of bargainers (as opposed to just revealing it), how periods of conflict are succeeded by periods of industrial peace (and vice versa), and how the stability of bargaining protocols depends not only on the conventions regulating the relations between trades unions and firms but also on those between workers and trades union leaders as well as on technological innovations.

5.2.2 Setting up the model

In any analysis of rational bargaining between trades unions and firms, industrial conflict must be explained as the result of some informational deficiency. For, if the two sides knew in advance the outcome, their rationality ought to instruct them to settle in accordance with the foreseen outcome without incurring the cost of fighting. There is nothing controversial in this. However, the seeds of controversy take root at the next level of abstraction when it is assumed that bargaining problems have uniquely rational solutions to be deduced logically.

The starting point of this section is a recognition that a complete range of rational bargaining strategies cannot be specified in advance even under perfect information about objective functions.

¹

Unlike the conventional literature, which seeks out equilibrium bargaining strategies after axiomatically imposing on firms and trades unions conjectures that are consistently aligned, an alternative evolutionary approach is suggested below. The objective is to explore the evolution of conventions which lead bargainers to aligned beliefs and, ultimately, to bargaining agreements.

As the above suggests, mainstream theory thinks of settlements between trades unions and firms as the realisation of uniquely rational strategies, and of the prospect of conflict (e.g. strikes, lockouts etc.) as the provider of information about the objectives and constraints of each other.

²

In this sense, the possibility of industrial conflict aids the revelation of the firm's and union's bargaining dispositions. Nevertheless, actual strikes cannot be accounted for unless they are blamed on some institutional constraint or on irrationality.

³

By contrast, my evolutionary approach sees automatic settlements as evidence that one out of many equally rational conventions has become established, and of conflict as both a byproduct of the process of convergence to a convention and as a symptom of the mutations which periodically threaten every such convention. In this context, industrial conflict plays a significant role in the creation of the bargainers' dispositions.

Section 5.2.3

establishes the notion of bargaining strategies as the products of evolution.

Section 5.2.4

then illustrates the new insights made possible by an evolutionary approach. In

particular, it suggests a new interpretation of strikes as experiments with alternative evolutionary protocols and illustrates how periodic waves of strike activity may be due to rational tests of the evolutionary stability of the status quo as well as to the separate conventions regulating the relationship between trades unions and their constituents.

Section 5.2.5

concludes.

5.2.3 Evolving bargaining strategies

Suppose a trades union (U) and a firm (F) have access to a history of H negotiations. This history can be thought of as a database or matrix with each of the H rows representing one negotiation while the columns of this matrix correspond to each bargaining round. Of the H rows (or negotiations), h rows involved the same U and F pair whereas the remaining $H - h$ relate the history of negotiations between other firms and unions in the industry or related industries. Each column contains a pair of demands (one for F the other for U) for each round of the negotiation. These pairs of U and F demands are expressed in terms of portions of the surplus to be distributed between capital and labour

4

which is normalised to equal 1 for convenience; i.e. $(x, y)_t$. If in negotiation i agreement was reached at, say, $t = 2$ then the entries for columns $t > 2$ are left empty.

To summarise, each negotiation is remembered by a string of demands

$$((x_1, y_1), (x_2, y_2), \dots, (x_\tau, y_\tau))_i \text{ s.t. } x_\tau = 1 - y_\tau$$

implying that agreement was reached in round τ . A negotiation i characterised by $\tau = 1$ is one which achieved agreement without a strike. Thus if $i \in H$, where H is the set of all previous negotiations, and C is the subset of H whose elements involve instances of industrial conflict, then $\tau > 1$ for all $i \in C$.

Imagine that the current negotiation is in round t . Of the H available observations, F samples m_F past negotiations which had also reached round t . Letting k_F be the number of observations out of m_F in which U accepted, during t , an offer equal to or less than $1 - x_t$, then F 's empirical cumulative distribution function of the probability that the union will accept $1 - x_t$ is $G(1 - x_t) = k_F/m_F$. If the trades union accepts this offer, then F 's payoffs in round t equal $U^F(x_t)$; otherwise it incurs the cost of an extra round of delay in reaching agreement, say c_F . For simplicity we assume that these conflict costs are constant, i.e. $c_F(t) = c_F(t - 1)$. Hence F 's per round optimal demand is given by (

5.1

):

$$x_{it}^* = \operatorname{argmax} \{U^F(x_{it}) + c_F(t)\} (k_F/m_F) \quad (5.1)$$

A similar description of U 's optimisation problem yields its optimal demand per round per negotiation as (

5.2

) below:

$$y_{it}^* = \operatorname{argmax} \{U^U(y_{it}) + c_U(t)\} (k_U/m_U) \quad (5.2)$$

Therefore observed strike duration τ in each negotiation is the minimum value of t which gives rise to (

5.3

):

$$\begin{aligned} &\operatorname{argmax} \{U^F(x_{it}) + c_F(t)\} (k_F/m_F) \\ &- \operatorname{argmax} \{U^U(y_{it}) + c_U(t)\} (k_U/m_U) = 0 \end{aligned} \quad (5.3)$$

At this stage it is worth noting the difference between an equilibrium and an evolutionary approach. The former tradition treats the probability of disagreement in each stage as a set of subjective beliefs of F and U to be worked out in a way such that (a) they are

consistently aligned (or common knowledge) inter-temporally and (b) they are consistent with (

5.1

) and (

5.2

) above. For this to be possible the implicit assumption is made that such a uniquely rational set of beliefs exists; Sugden (1990) calls this the *axiom of rational determinacy*. By contrast, the evolutionary approach has agents accepting the impossibility of such an a priori coincidence of beliefs. Once they recognise the plausibility of many alternative subjective beliefs about each other, they look to past experience for a guide to the negotiation in hand.

Notice that this is not to say that bargainers opt for adaptive learning because they are less than rational; it is rather that rationality cannot pick out the 'right' beliefs and therefore bargainers' only real option is to blunder around for clues, acting as sensibly as they can. Perhaps the most striking difference between the evolutionary and the conventional equilibrium approach is that the former attempts to generate endogenously the equilibrating mechanism whereas the latter imposes it axiomatically. The fact that the conventional literature has only provided a thin explanation of rational strikes (see previous chapters for proof) is no more than a natural reflection of the methodological move to assume, as opposed to generate, equilibration of beliefs. Once beliefs are assumed to be in alignment, it is unsurprising that the only explanation of failing to avert costly disagreement (even under asymmetrically distributed information) is either some exogenous impediment to settling quickly or irrationality – see footnote 3. The promise of the evolutionary approach is that strikes can be admitted as the result of rational behaviour by agents who are searching for a way to equilibrate their beliefs.

To offer an idea of how this equilibration can occur endogenously, consider a strike which has been going on for t rounds already. What offer should F and U make at $t + 1$? Judging from (

5.1

) and (

5.2

), it seems that the answer depends on the number of times in the past that particular offers under consideration were accepted by the opposite side divided by the number of negotiations that also lasted $t + 1$ rounds. However, notice that the longer the strike the fewer the observations m_F and m_U which are left into the sample. So, if bargainers were to base their estimate of the cumulative probability distribution function of having an offer rejected at $t + 1$ solely on the empirical equivalent (i.e. on the ratios k_j/m_j , $j = U, F$) they would effectively be rejecting valuable information. For example, suppose that in round $t + 1$ U is looking at a previous negotiation which was settled in round t with F accepting U 's demand of, say, y' . That negotiation

never reached round $t + 1$. Does this mean that it should drop out of U 's current sample? Does it not contain useful information on whether F may accept U 's demand for y' at $t + 1$?

However, if this observation of what happened is admitted in the sample, it will increase k_U and m_U by one, effectively increasing the estimated probability that an offer to F at time $t + 1$ of $1 - y'$ will be accepted. While it is difficult to argue that U should not change its prediction that F will accept $1 - y'$ in this way (since the fact that $1 - y'$ was accepted in a previous negotiation in a similar, albeit earlier, round carries interesting information), on the other hand such an alteration of the available sample is largely arbitrary: U does not have any firm indication of how F would have behaved in round $t + 1$ of the previous negotiation since that round was never reached. Whether a bargaining side will proceed with this alteration or not (and in the absence of a uniquely

rational bargaining strategy), is a matter of disposition. Some trades unions or firms may admit this type of deduction in their information set, whereas others will not.

Let d_t^F and d_t^U denote the two sides' dispositions in this regard, defined as the number of previous negotiations which will be sampled in order to gauge what will happen in the next round (i.e. $t + 1$) even though they were settled in some round $t' < t + 1$. Thus the optimal offers in (

5.1

) and (

5.2

) become:

$$x_{it}^* = \operatorname{argmax} \{U^F(x_{it}) + c_F(t)\} [k_F + d_t^F] / (m_F + d_t^F) \quad (5.4)$$

$$y_{it}^* = \operatorname{argmax} \{U^U(y_{it}) + c_U(t)\} [k_U + d_t^U] / (m_U + d_t^U) \quad (5.5)$$

Expressions (

5.4

) and (

5.5

) give the evolutionary bargaining process its foothold. Since the parties' dispositions are arbitrary, it is they that must evolve through time in response to aggregate behaviour. Letting $D^U \times D^F$ be the set of all possible dispositions, we think of $\rho_{it}(d^U, d^F)$ as the probability that during round t of negotiation i the set of dispositions (d^U, d^F) will be selected by the two sides out of set $D^U \times D^F$. The question then becomes: how will these dispositions evolve?

Given a history $h < H$ between F and U and a particular set of dispositions, in each round of the current negotiation (

5.4

) and (

5.5

) translate into each negotiating team's optimal mixed bargaining strategies:

$q_F(x | d^F, h)$ is the conditional probability that F offers U share $1 - x$ given history H

$q_U(y | d^U, h)$ is the conditional probability that U offers F share y given history H

Let us now assume that q_F and q_U are best-reply probability distributions such that $q_F(x | d^F, h) > 0$ only if x happens to be a best response by F to the sample drawn from history H given its disposition. Similarly, $q_U(y | d^U, h)$ is presumed to be U 's best reply probability distribution of bargaining strategies in a particular round given the available information and U 's disposition.

In order to illustrate the evolutionary mechanism, we follow the standard method of inquiring about the possibility of the bargaining process reaching a stationary state. The aim is to show under what conditions the bargaining history

between F and U may become sufficiently stable in order to explain the equilibration of beliefs (i.e. q_F and q_U tending towards $E^F(q_U)$ and $E^U(q_F)$ respectively). Let h' be an alternative history of round per round negotiations between F and U (equal in size to h , the number of negotiations that has already involved F and U). We call h' a successor of h if there is a non-zero transition probability ($R_{hh'}$) that h' will follow immediately after h .

$$R_{hh'} = \sum_{f \in D_F} \sum_{u \in D_U} \rho(d^F, d^U) q_F(x_t | d^F, h) q_U(y_t | d^U, h) \quad (5.6)$$

DEFINITION 5.1 Let us define a set of offers $(x^*, y^*)_t$ as a bargaining convention if it denotes agreement (i.e. $x^* = 1 - y^*$) and has occurred in the same round of h^* successive negotiations. [Notice that if such a bargaining convention is realised, and provided h^* is sizeable enough, the particular choice of sample (i.e. the bargaining

dispositions d_F and d_U will no longer affect behaviour).]

PROPOSITION 5.1 *Once a convention is established, industrial conflict vanishes.*

Proof: A convention marks an absorbing state of the generalised bargaining process described by the transition mechanism in (

5.6

). Since probabilities q_F and q_U [see (

5.4

) and (

5.5

)] are assumed to be best replies to the available information, the best reply to a history of h^* successive $(x^*, 1 - x^*)_t$ agreements in round t of each negotiation is for F to offer and for U to demand $1 - x^*$ in round t . But then as long as the costs of disagreement (c_F, c_U) are positive, and through a process of backward induction, it transpires that τ tends to 1 as the bargaining process in (

5.6

) approaches an absorbing barrier. \square

ASSUMPTION 5.1 *If bargainers have a disposition to seek information about the current round t in d^F and d^U past negotiations which ended in a number of rounds less than t , then they look at the most recent d_F and d_U negotiations from the available record of H negotiations.*

PROPOSITION 5.2 *If at least one bargaining disposition ($d^U \leq D_U$ or $d^F \leq D_F$) chooses a sample of at most half of existing records (i.e. $d^F, d^U \leq H/2$), then from any initial state the bargaining process will converge to a convention with high probability in a finite number of negotiations.*

The above proposition is an extension of the first theorem of Young (1993) which applies to a series of Nash games played once by pairs randomly drawn from a fixed population. By showing that a convention is most likely to emerge it endogenises the equilibration of bargainers' beliefs. For, if the bargaining process can be shown (as opposed to being assumed) to generate a single agreement as time goes by, then it is plausible to expect rational bargainers to align their expectations. The central difference between this result and the conventional Rubinstein-based solutions (see his 1985 paper, and

Chapters 3

and

4

) is that the

point of agreement $(x^*, 1 - x^*)$ is one of many equally plausible outcomes and could have easily been otherwise (i.e. unlike Nash and Rubinstein, the evolutionary model herein does not assume that the evolved settlement reflects a uniquely rational bargaining solution).

In summary,

Propositions 5.1

and

5.2

suggest that the process of negotiations, rooted in its own history, founders for a while until it generates a convention. 'Foundering,' in this context, translates into industrial disputes. Once a convention is in place, trades unions and firms manage to coordinate their beliefs in accordance to the established convention. The difference with equilibrium theory is that our approach appreciates the impossibility of determining theoretically which convention will emerge.

Sketch of proof: The aim is to show that the set of all strategy choice paths which do not lead to an absorbing state (that is, a convention) has a vanishing probability. To do

this I shall prove that there exists an integer I and a positive probability ξ such that the probability of converging to a convention within αI ($\alpha > 0$) negotiations is $1 - (1 - \xi)^\alpha$. For if this is so, then as $\alpha \rightarrow \infty$ the probability that a convention will be reached will tend to one. Thus the proof that a convention will be reached within a finite number of negotiations. The formal proof is located in the appendix to this chapter.

Summarising

Proposition 5.2

, there exists a positive probability that a convention can be reached within a finite number of negotiations. Hence, there exists a positive stationary probability (that is, independent of the particular history) that the history of negotiations will engender some convention which allows for agreements without industrial conflict.

5.2.4 Strikes as experiments with alternative conventions

Conventions are genuinely absorbing states to the extent that bargainers consistently choose demands as best replies to the demands of their opponents. By its very nature, a convention makes sense to each firm or trades union when others also subscribe to that convention. However, this does not mean that a current convention is in the interest of each party, or indeed of a majority of trades unions or firms *even if it helps them avoid costly strikes*. The reason is that, as evolutionary game theory shows [see

Chapter 6

in Hargreaves-Heap and Varoufakis (2004)], the evolutionary fitness of conventions is increased when they treat different types of agents in different ways. For example, a convention may give a trades union $1 - x^* = 1/2$ when it is bargaining with a firm located in the manufacturing sector but only $1 - x^* = 1/4$ when bargaining in the service sector. The point here is that the emergence of the convention will benefit the average trades union (or indeed firm) but if some union happens to be so placed with respect to the convention that it gets the richer rewards infrequently (e.g. because its members are located mostly in the service sector), perhaps it would be better off without the current convention.

One is justified to ask: Why does the trades union then stick to the convention, if it would be better off without it? The answer is that even though the individual

trades union would be better off if all bargaining parties were to abandon the convention, it does not necessarily make sense to do so individually. For example, it could simply trigger a much longer strike to get something above $1 - x^*$ simply because the firm's expectations are fixed on the focal point provided by the convention. However, the extent to which a convention has the capacity to reproduce itself, and therefore to thwart such attempts to re-write the evolved bargaining protocol, depends on the degree to which a critical mass of bargaining units in the labour market are willing to risk some industrial conflict in order to test the stability of a particular convention. This inquisitiveness of agents is what marks them apart from the purely adaptive automata which the rational expectations hypothesis was meant to sideline.

To make the last point more sharply, in conventional equilibrium (i.e. neoclassical) theory the urge to see ahead, and to avoid becoming bogged down in an equilibrium whose only support comes from the past, takes the form of a rational expectation. Rational expectations are then derived by postulating a correct model of expectation formation and subsequently allowing bargainers access to it. However, this presumes that a 'correct' model can be specified in advance based solely on information concerning objective functions and constraints. In an evolutionary framework, however, the possibility of such fore-knowledge of the 'correct' model is rejected in view of the multiplicity of equally plausible candidates (i.e. of 'correct' models) out of which one materialises in a radically unpredictable manner (see

Propositions 5.1

and

5.2

).

In this framework forward-looking agents recognise that the current convention is characterised by different degrees of stability which depend on aggregate behaviour. In the absence of uniquely rational expectations about the evolutionary stability of this convention, they do the one thing that rational agents can do: they experiment by testing the effect of their individual industrial action on aggregate bargaining behaviour. For instance, an established convention may award $1 - x = 0.6$ to workers in the construction industry and only $1 - x = 0.2$ to miners. The mining unions know that, if they abandon the convention (which has them accepting 0.2 without a strike) a strike will follow.

Whether they will benefit from it depends on whether their action at $t = 1$ will cast sufficient doubt in the mind of employers at $t = 2$ as to whether their optimisation calculations, based on the current convention, are still valid. It will also depend on whether trades unions in other industries, who have also been doing less well as a result of the current convention, are prepared for industrial conflict. A similar story can be told about employers who decide to test the stability of a convention which discriminates against them in favour of firms in other industries. In this context, industrial conflict suddenly emerges as the byproduct of experimentation. And, unlike neoclassical theory's interpretation of conflict as a mere provider of information about exogenous types of bargaining behaviour, evolutionary theory argues that conflict helps create the prevalent types of bargaining conduct.

The next question which needs to be addressed concerns the precise form of these experiments. We discuss two types: (a) Strikes which reflect random experiments, and (b) strikes due to experiments which are causally related to some underlying historical, technological or political process.

Random, uncorrelated experiments

Imagine that firms and trades unions test the stability of the current convention at random, hoping that they can re-jig it in a manner which boosts their returns. Let θ_F and θ_U be the probabilities with which F and U respectively would experiment in any given round of the negotiations, and $Q_F(x | d^F, h)$ and $Q_U(y | d^U, h)$ be the replies of F and U to their observations of the past when they decide to experiment. Then the transition probability from one history (h) to another (h') becomes:

$$\begin{aligned} R_{hh'}^\theta = & \sum_{f \in D_F} \sum_{u \in D_U} \rho(d^F, d^U) \{ (1 - \theta_F)(1 - \theta_U) q_F(x_t | d^F, h) q_U(y_t | d^U, h) \\ & + (\theta_F)(1 - \theta_U) Q_F(x_t | d^F, h) q_U(y_t | d^U, h) \\ & + (\theta_U)(1 - \theta_F) q_F(x_t | d^F, h) Q_U(y_t | d^U, h) \} \\ & + (\theta_F)(\theta_U) Q_F(x_t | d^F, h) Q_U(y_t | d^U, h) \end{aligned} \quad (5.7)$$

When the θ s are uncorrelated with each other or across different negotiations, the bargaining process may still gravitate towards a state of (mostly) industrial peace but will be punctuated with the odd strike. An occasional random build-up of experimental deviations may snowball into a chain reaction of industrial unrest which will again die down provided the variance of the θ s is not too high.

5

Consider the convention towards which bargaining outcomes would have gravitated in the absence of random experiments (or strikes). Will it survive? Or will another distribution of the surplus between workers and employers become the new attractor of bargaining processes? The answer depends on the stochastic stability of the initial convention. Some will prove more resilient than others.

In technical terms, a convention h^* is stochastically stable if $RR_{hh'}^\theta$ has a unique stationary distribution according to which the bargaining process proceeds as the magnitude of the experiments vanishes. In that case, the probability that the distribution of the surplus will be determined by convention h^* exceeds at any stage the probability that it can be better explained by any other convention.

Historically correlated experiments: the effect of technological innovation and trades union politics on the probability of experimentation

Although an interesting history of industrial relations has been made possible without having to ascribe experimentation to anything other than rational curiosity (symbolised by random disturbances), the present approach allows more to be said on the determinants of such tendencies. I examine two cases. The first refers to technological innovations which alter the costs of conflict. Suppose, for example, that a convention has evolved such that a trades union and a firm habitually settle on $(x^*, 1 - x^*)$ without conflict. Suddenly, some technological innovation alters the production process in ways which affect the firm's objective function and/or conflict costs. For instance, if the new technology renders redundant middle-ranking supervisors loyal to the union, the trades union will have lost a major weapon with which to inflict costs on the firm (e.g. in terms of shutting down production quickly). This development, by itself, may be sufficient to destabilise the convention and to give rise to a period of conflict before some other convention unfolds. The UK print media in the 1980s offered a suggestive example.

The second case considers the effect of workers' expectations on the trades union leaders' propensity to subvert the existing convention. Noting that such a decision can only make sense provided the union's members are prepared to back their leaders' recalcitrance by walking out, it is interesting to explore the linkages between the 'experiments' with alternative conventions and the workers' beliefs. Consider the first round of some negotiation. Probability θ_U relates the chance that the trades union will breach the prevailing convention $(x^*, 1 - x^*)$ by rejecting the firm's $1 - x^*$ offer in round $t = 1$. Instead it demands $1 - x'$ in round $t = 2$, where $x' < x^*$. In this case, $z = x^* - x'$ is the extent to which the trades union aims to alter the portion of the surplus which has so far been retained by the firm conventionally.

For the purpose of illustrating the new analytical possibilities, let us suppose that union leaders care about what workers expect concerning their tactics – especially if the latter involve strike calls whose success will depend entirely on how workers respond to them. Also, workers may evaluate their leaders' tactics according to what expectations they have of them. Workers, for example, may prefer their union to breach a convention and to struggle for the establishment of a more beneficial distribution of the surplus if, for some reason, this is what they expect the union to do. And conversely, they may be disappointed if the union calls for a strike which they had not anticipated. The above suggests an intricate web of beliefs which may constitute an important part of what keeps the trades union a viable organisation in the face of all sorts of prisoners' dilemmas.

To extract from the above an analytical contribution, let θ' be the workers' estimate of θ_U and θ'' the union's estimate of θ' : $\theta' = E_{\text{workers}}(\theta_U)$ and $\theta'' = E_{\text{union}}(\theta')$.

Table 5.1

offers an analytical counterpart of the above paragraph.

Note that the payoffs are arbitrary and only hope to illustrate the relative effect on the leaders' and workers' utility following the decision of the former to abide by, or to disregard, an already established rule (i.e. convention) for splitting the firm's surplus between capital and labour.

If workers expect a deviation from the convention (and thus a strike) with a high probability [$\theta' > 1/(1 + z)$], then they prefer their leaders to deviate from the convention and call a strike. If they are not so sure that a deviation is as likely, then they will not be disappointed if their trades union respects the convention and

settles immediately. In this example, what matters most is that workers' expectations of their leaders' bargaining tactics are confirmed. The interesting twist here is that, if union leaders think that their constituency expects them to deviate, then they *want* to deviate. If not, they feel no need to break with the convention. They may still do so with positive probability, e.g. $\theta_U = \eta$; $\eta: N(0, \sigma^2)$, as part of the usual experimentation with alternative conventions, but they will not introduce a systematic disturbance into the bargaining process of (

5.7

).

Let us consider the following condition which must be satisfied for the continuation of a largely strike-free period once a convention has been established: $\theta_U = \theta' = \theta'' = 0$ – that is, no deviation is planned by leaders, none is expected by the workers and, finally, leaders do not feel they are expected to deviate. Notice that this outcome yields the highest possible payoff for both workers and leaders, viz. the collective attitude towards the convention. Interestingly, this does not mean that the convention is necessarily safe. Consider two possibilities:

Firstly, some political developments in the industry or elsewhere may generate in workers' minds the idea that the trades union is about to, or should, deviate from the convention and thus cause a strike. Then the leaders will be trapped in the workers' expectations which, in a never-ending circle, they will have an incentive to confirm even though they are perfectly aware of the fact that this alternative equilibrium of beliefs ($\theta_U = \theta' = \theta'' = 0$) yields a lower payoff for all involved.

Secondly, leaders may conclude that the prevailing convention is unstable and that a reasonably intense period of industrial unrest will bring into being a far more propitious distribution of the surplus. They embark upon a political campaign whose purpose is to prepare the workers for the deviation. Once $\theta' > 1/(1 + z)$ they are free to deviate and reject the firm's offer at $t = 1$. Underlying this argument is the thought that a union leader preparing for a strike will want workers to approve the 'deviation'. But as

Table 5.1

reveals, all that may be required is that workers are cajoled into expecting a deviation. Once the political campaign achieves this, a deviation follows naturally.

There are two lessons from this: First, the tendency to deviate from a convention (and thus to rekindle social and industrial conflict) may be, to a significant extent, socially and politically determined. Trades union leaders are neither mere conduits for workers' preferences, nor unscrupulous purveyors of self-serving tactical manoeuvres. Similarly, workers are neither passive playthings of the trades union's internal politics, nor sovereign creators of bargaining strategies. Secondly, the fact that a particular convention may seem safe, because its continuation

receives support from Pareto-dominance, does not mean that rational trades unions (and indeed firms) should not attempt to subvert it.

9

Industrial conflict suddenly becomes a much richer social phenomenon than the conventional neoclassical view of it permits.

Table 5.1

When the game's structure depends on second-order beliefs

	<i>Union leaders' utility from choice of z and θ_U ceteris paribus</i>	<i>Workers utility from choice of z and θ ceteris paribus</i>
Leaders choose to deviate from	θ''	θ'

convention h^* by z		
Leaders accept convention h^*	$1 - (\theta''/z)$	$1 - (\theta'/z)$

5.2.5 Conclusion

J. R. Hicks (1966) was not entirely wrong when he famously suggested that ‘... most strikes are the result of faulty negotiations’. The truth of his statement hinges on the interpretation of these ‘faults’ or ‘errors’. If one assumes, as neoclassical bargaining theory does, that there exists a model of uniquely rational strategies, then ‘errors’ can be avoided by adopting this model and strikes happen when people are not rational enough to do so.

In contradistinction, if one believes, as the evolutionary perspective here recommends, that no such model can be worked out a priori, then what appear as negotiating ‘faults’ are the necessary steps rational bargainers must take to defeat the unavoidable indeterminacy of bargaining. Strikes are the symptom of these failed attempts along the evolutionary path to stability. They are also a symptom of the arbitrariness of any convention which opens it up to frequent challenges. Those challenges are not only a result of the rational inquisitiveness of trades unions and employers alike, but are a reflection of the instability of other underlying conventions as well (e.g. those governing the internal politics of a trades union).

All this translates into a rich history of continually established and subverted rules according to which a firm’s surplus is distributed between capital and labour. Strikes are the natural symptom of the evolution of this distribution and a rational outcome of an irrepressible indeterminacy aided and abetted by human’s capacity to reason in creative ways that the neoclassical mind refuses to fathom.

5.3 Rational rules of thumb in finite dynamic games

5.3.1 Introduction: toward a model of *N*-person backward induction with inconsistently aligned beliefs and full rationality

Let us now return to the end of

[Chapter 4](#)

, where I concluded that backward induction, when combined with the enforced equilibration of beliefs (the strict version of the neoclassicists’ *third meta-axiom*, see [Chapter 1](#)

), resolves indeterminacy but only at the hefty cost of logical incoherence. The point of my

[Chapter 4](#)

critique of the neoclassical approach is that the resulting consistently aligned beliefs (CAB) are incoherent in view of the counterfactuals they rely on. The current section also focuses on the infamous centipede game (as did

[Chapter 4](#)

, see

[Figure 4.1](#)

). It does so with a view to examining what rational people might do when unshackled from the logical incoherence of the neoclassical analytical method. It asks: How will the possibility of inconsistently aligned beliefs affect the manner in which rational players play such games? It shows that, provided beliefs are aligned monotonically, some of the interesting qualitative features of the conventional approach remain unchanged while, at the same time, a much richer behavioural pattern is recognised and the logical incoherence of neoclassical game theory is done away with.

In short, this section takes its cue from an earlier conclusion that the neoclassical analysis of finite dynamic games is incoherent (and thus misleading), and it investigates

what we can say (once we reject the neoclassical approach) about how rational players would behave in the context of centipede-like games. To keep the analysis as general as possible, one of the following subsections generalises to versions of the game involving more than two players.

5.3.2 The centipede game in 'coin' form

Rather than the standard depiction of the centipede game that we used in

Figure 4.1

(see the previous chapter), it is often more inspiring to couch it in terms of a game involving coins. Imagine a table piled up with G gold sovereigns. Two or more players take turns to collect either one or two coins at a time. If the active player collects one coin, then the next player gets a chance to do the same. If on the other hand she collects two coins, the game ends. For this reason taking one coin (or playing ACROSS) will be thought of as a 'cooperative' move, thus labelling the taking of two coins (playing DOWN) a 'defection'.

Figure 5.1

offers the extensive form representation of the game (which is analytically indistinguishable to that of earlier

Figure 4.1

) which points to the usual paradoxical 'solution:' Under the composite assumption that players' beliefs are (a) formed by backward induction and (b) are subject to common knowledge of instrumental rationality (CKR), the game ends immediately with the first player taking two coins.

That this conclusion is paradoxical there is no doubt: Firstly, experimental evidence does not support it.

10

Secondly, it does not get easier to accept the more intelligently we think about it (especially if G is large). Indeed, there have been a number of philosophical and logical objections to the legitimacy of imposing (a) and (b) above simultaneously; an analytical move tantamount to assuming that agents invariably entertain CAB.

11

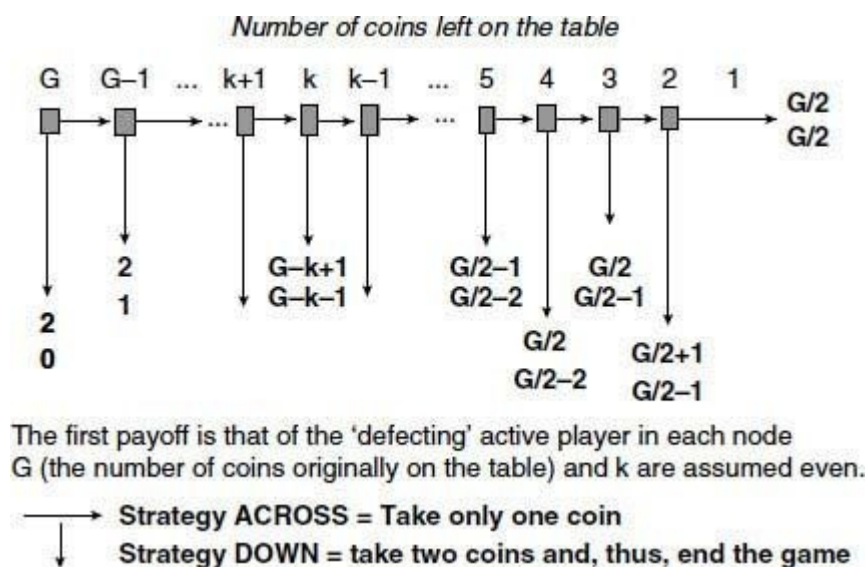


Figure 5.1

The centipede game revisited.

The question now becomes: How will rational players, who recognise the illegitimacy of the CAB assumption, play this game? Will they act in a manner qualitatively different to that prescribed by models which retain CAB after introducing uncertainty about the rationality of one's opponent? The conclusion is that, provided beliefs are aligned monotonically (albeit inconsistently), we can retain the more interesting features of the latter without taking steps (such as assuming CAB) which are difficult to defend on

philosophical grounds. However, there is a price to pay, as the solution depends on an arbitrary choice regarding the degree of alignment of people's beliefs. Nevertheless, this might be inevitable since the major point of the

Chapter 4

critique was precisely that strategic behaviour yields inherently unpredictable degrees of belief alignment.

5.3.3 The two-person version

Backward induction, together with CKR, leads to the robust conclusion that no instrumentally rational player will ever take just one coin. Yet the paradox here is that, in order to work out why, one needs to consider what will happen at the last stage of the game first, then at the penultimate stage... and so on; that is, it must be pre-supposed that players have chosen only one coin many times already. It is clear that, if their beliefs were consistently aligned, the game would not have moved into these later stages. But, in order to work out these beliefs we need to consider these stages; a messy sequence of counterfactuals which can only be tamed provided we are prepared to assume that agents go that far into the game as a result of random mistakes (often referred to as 'trembles') which occur (i) with tiny probability, (ii) are independent of agents' beliefs, and (iii) remain uncorrelated across stages.

Evidently, the longer the game the greater the amount of 'trembling' people must consider as probable before they work out (backwards) what it is rational to believe at the outset and, thus, the less convincing the theory. Additionally, the more coins there are on the table the more difficult it is for instrumentally rational players to discern the difference between 'trembles' and bluffs – recall the last chapter once again. Having recognised these difficulties with CAB, let us begin with a question:

'Why would an instrumentally rational player ever choose only one coin when it is her turn to play?' Answer: 'Only if she had rational grounds to expect that the next player will also choose a coin with a probability of at least 1/2' Consider the stage with k coins left on the table at which player A^k is active and let:

$p_k = Pr(A^k \text{ will choose 1 coin})$ (i.e. play ACROSS)

$\pi_k = Pr(A^k \text{ is motivated by non-instrumental reasons})$

$q_k = Pr(A^k \text{ is motivated by instrumental reasons but will still choose only 1 coin})$

An instrumentally rational player chooses in a manner that maximises her payoffs given the rules of the game and her beliefs about the other player. Hence an instrumental player will always take two coins when there are three left on the table. However, she may resist the temptation and pick up only one coin if there are $k(>3)$ coins left and she expects her opponent also to take a single coin during the next stage.

This is what I mean by an 'instrumental reason' for choosing one rather than two coins at k . By contrast a player who is motivated differently (e.g. is concerned with fairness, or has adopted some universalisable principle of practical reason, or follows a social convention of sharing etc.) is assumed *always* to choose one coin. This assumption could, of course, be relaxed by introducing an exogenous probability with which a non-instrumental player chooses two coins. For simplicity, we assume that this probability is zero.

Let us focus on an instrumental player at stage $k + 1$. For the game to have reached k , this means that A^{k+1} set p_k to be greater than 1/2.

Equation (5.8)

captures her expectation:

$$E_A^{k+1}(p_k) = E_A^{k+1}(\pi_k) + [1 - E_A^{k+1}(\pi_k)]q_k \quad \text{and}$$

$$q_k = E_A^{k+1}[Pr(E_A^k(p_{k-1}) > 1/2)] \dots \quad (5.8)$$

where $E_A^k(\cdot)$ denotes the expectation of player A^k who is acting (and thinking) instrumentally. The assumption of CAB imposes the following equilibration:

$$E_A^{k+1}[Pr(E_A^k(p_{k-1}) > 1/2)] = E_A^k[Pr(E_A^k(p_{k-1}) > 1/2)] \quad (5.9i)$$

$$E_A^{k+1}(\pi_k) = E_A^k(\pi_k) \quad (5.9ii)$$

It is easy to see how, under backward induction, the above two conditions mean one of two things: Either $q_i = 0$ for $i = 2, 3, \dots, G - 1$, which means that the instrumentally rational player who opens the game (i.e. player A^{G-1}) does not expect the other to take with probability more than 1/2 only a single coin when there are 3 left, i.e.

$E_A^{G-1}(\pi_3) < 1/2$. Or, if $E_A^{G-1}(\pi_3) > 1/2$, $q_i (> 0)$ is common knowledge (given CAB)

and is computed by means of Bayes' rule backwards.

12

Let us now consider the case in which players do not trust that the conditions for CAB [

5.9(i)

and

5.9(ii)

] should be taken for granted. As an example, consider first the stage where $k = 3$. Clearly, $q_3 = 0$ and therefore $p_3 = \pi_3$. At stage $k = 5$, p_5 will exceed π_5 provided $q_5 > 0$. Would it be rational for player A^5 to entertain such an expectation? The moment we are prepared to accept the possibility that rational players got to stage $k = 5$ without assuming that they did so as a result of uncorrelated, independently and identically distributed random errors (that is, as long as we allow for the possibility of inconsistently aligned beliefs), then it is inevitable that $q_5 > 0$. Thus, it turns out that when there are 5 coins on the table the probability of a 'cooperative' move is greater than at the later stage when there

are three coins left ($p_5 > p_3$). If by symmetry $q_7 > q_5$ then, from

equation (5.8)

it transpires that $p_7 > p_5$. And so on.

In effect, we have come to an important conclusion without any controversial assumptions: Since the propensity of rational players to pick up a single coin (when it is their turn to play) is an increasing function of the expectation on the left hand side of

equation (5.8)

, the more the coins on the table the more likely that the instrumentally rational player will 'cooperate'.

To take this observation further, three basic assumptions are required:

- Symmetry – $S: p_k = E_A^{k+1}[p_k | \Upsilon(A^{k+1})] \forall k$, where $\Upsilon(A^{k+1})$ is the information/belief set of player A^{k+1}

i.e. instrumentally rational agents will play cooperatively with a probability that others like them would have estimated in an unbiased manner had they had access to their beliefs. This assumption allows for beliefs to be inconsistently aligned (since q_k is not known with certainty to player A^{k+1}) yet demands that players have the same computational capacities and thus, makes it possible to trace the path of p_k given assumptions R and M below.

- Instrumental reflection on non-instrumental agents – $R: \pi_k > 0$ and $\Delta^2 \pi_k > 0 \square k$
i.e. there is always a possibility the next player will choose to take a single coin non-instrumentally. Moreover, the chances of this happening cannot decrease with the number of coins left on the table. In the simplest case ($\Delta^2 \pi_k = 0$), this probability is constant and corresponds to the proportion of (non-instrumentally) cooperative persons in the population. In the more general case, instrumental agents reflect that the larger the number of coins left the greater the possibility that normative expectations favouring cooperation will emerge which cannot be explained instrumentally.

- Monotonically aligned beliefs – $M:$

$$q_k = \{E_A^{k+1}[Pr(E_A^k(p_{k-1}) > 1/2)] = f(p_{k-1}) \text{ where } f'(\cdot) > 0 \quad (5.10)$$

Condition (

5.10

) replaces (

5.9i

). Whereas (

5.9i

) imposes a strict equality between the beliefs of player A^{k+1} and of A^k , viz. the chances that A^k will expect a cooperative move at stage $k - 1$, condition (

5.10)

issues the far less stringent (and therefore defensible) requirement that their beliefs are linked monotonically. This is equivalent to the thought that, if one is attempting to assess the probability, say γ , of another person predicting that some other probability, say δ , exceeds $1/2$, then it is reasonable to expect that γ will be an increasing function of δ . Clearly, this assumption imposes *some* alignment between players' beliefs without going to the extremes of the neoclassical CAB axiom. How much alignment there will be, of course, depends on the precise functional form of $f(\cdot)$. The point of the critical literature on the question of alignment (see

Chapter 4

) is that due to the inherent unpredictability of human nature, there exists no unique $f(\cdot)$, i.e. one derivable in a uniquely rational manner.

The repercussion of the three assumptions, S , R and M , above is simple:

equation (5.8)

reduces to the difference equation

$$p_k = \pi_k + (1 - \pi_k) f(p_{k-1}) \quad (5.11)$$

Given some idea about the form of $f(\cdot)$ and the probability that a player will cooperate for non-instrumental reasons when there are k coins on the table, we can trace the path of the probabilities of cooperative moves by instrumental players. A similar, yet independent, sequence can be found for q_k .

5.3.4 Generalising for N players

With N players taking turns to collect their one or two coins from the table, it is clear that cooperation requires either a large number of coins or a smaller short-term advantage from defection. To extend the analysis so that it applies to a range of payoffs, suppose the rules specify that a player whose turn it is to act can collect either D or C coins, where $D > C$ (that is, D corresponds to the defection strategy and C to the cooperative move). So far in our game $D = 2$ and $C = 1$.

Let $d = (D - C)/D$. Then

equation (5.11)

generalises to (

5.12

) below in which, again, p_k is the probability assessment by player A^k of the likelihood of player A^{k-1} cooperating (if given a chance):

$$p_k = \pi_k + (1 - \pi_k) f\left(\prod_{m=2}^M p_{k-m'} d\right) \quad (5.12)$$

where $M = N$ if $k - N \geq 1/(1 - d)$ and $M = k - 2$ otherwise; and $f(y, d)$ is the probability with which A^k expects the next player to expect y to be greater than or equal to d , given that she expects it to equal y .

Clearly the condition for an instrumentally rational player A^k to cooperate by choosing C coins (when there are k left on the table) is given by (

5.13i

) while the initial condition of difference

equation (5.12)

is in (5.13ii):

$$\prod_{m=0}^M p_{k-m} \geq d \quad (5.13i)$$

$$p_k = \pi_k \quad \forall k \in \left[\frac{1}{1-d}, N + \frac{1}{1-d} \right] \quad (5.13ii)$$

Naturally the way in which players' beliefs are aligned – i.e. function $f(\cdot)$ – determines the value of (

5.12

). Even though it is a premise of this chapter that a unique $f(\cdot)$ ought not be imposed, it is interesting to explore different specifications. Consider those implying that A^k will be certain (or totally undecided) of the next $k - N$ players' decision only if she were totally certain (or undecided) herself if in their position; i.e. $f(0, d) = 0$; $f(1, d) = 1$ and $f(d, d) = 1/2$. It is easy to show that, under these restrictions, cooperative moves are likely by instrumentally rational agents.

Table 5.1

reports on the minimum number of coins that must be left on the table in the two-person game before an instrumentally rational player cooperates (i.e. for condition (6i) to apply). The numbers correspond to the simple case where $f(y, d) = y/2d$.

Table 5.2

$N = 2$, $f(y, d) = y/2d$ (see also

Appendix 5.2

at the end of this chapter)

$\pi_j \backslash d$	0.2	0.3	0.4	0.5
0.1	5	6	8	10
0.01	12	22	36	55
0.001	84	184	324	505
0.0001	804	1804	3204	5005

Table 5.3

$N = 3$, $f(y, d) = y/2d$; empty cells denote that for cooperation to emerge the number of coins on the table must be infinite (see also

Appendix 5.2

at the end of this chapter)

$\pi_j \backslash d$	0.025	0.05	0.1	0.2
0.1	6	7	10	58
0.01	14	24		
0.001	88			

Table 5.4

$N = 3$; $f(y, d) = \Phi(\alpha + \beta(y - d))$ where $\Phi(\cdot)$ is a linear probit and (α, β) are chosen in a way that $f(y, d)$ is never one standard deviation away from the $f(y, d)$ function used in

Table 5.3

. Different choices for parameters α and β correspond to different assumptions about the degree of belief alignment between the three players (see also

Appendix 5.2

at the end of this chapter)

$\pi_j \backslash d$	0.025	0.05	0.1	0.2
----------------------	-------	------	-----	-----

0.1	5–6	6–7	9–11	56–60
0.01	13–15	23–25		
0.001	87–93			

5.3.5 Conclusion

Neoclassical game theoretical models of finite dynamic games are founded on particular assumptions which specify detailed stories about the players' out-of-equilibrium beliefs [in which normal form mistakes (or trembles) are introduced, i.e. trembles which are perfectly correlated across information sets]. For instance, Kreps *et al.* (1982) preserve a rigid structure of uncorrelated trembles while

introducing more than one type of player, each with a specific probability. In sharp contrast to this type of approach, the model above is based on very mild assumptions. Indeed, its starting point is the recognition that, in dynamic games, it is neither desirable nor possible to have detailed stories about how deviations from the equilibrium path are to be interpreted by players. Its conclusion is that, in addition to being theoretically undesirable, such detailed stories/assumptions are not even necessary. Moreover, the analysis offered herein, having accepted the inevitability of at least some inconsistency in the beliefs of rational agents, seems to be more in tune with the most recent results from controlled laboratory experiments (see Binmore *et al.*, 2002).

The reason why stringent assumptions about beliefs are undesirable is that pre-specifying particular patterns of trembles is incompatible with instrumental rationality in view of the counter-factual logic inherent in inducing beliefs via backward induction. On the positive side, the message here is that, even without such detailed stories, the important qualitative results usually derived from restrictive (and thus controversial) assumptions can survive without them. By making only minimalist assumptions (e.g. that people's beliefs are aligned monotonically, rather than consistently), we can still generate the same intuitively appealing predictions as those generated by means of the logically indefensible assumptions postulated by neoclassical economics. For instance, we find that the probability of a cooperative move by an instrumentally rational player:

- a. increases with the number of potential future stages (i.e. coins on the table);
- b. decreases as the number of players rises;
- c. rises with the expectation that agents may be motivated differently; and
- d. is inversely proportional to the gap between the payoffs from defection and cooperation.

5.4 Epilogue

This chapter is different from previous ones. It tries to go beyond criticism of the neoclassical method. Taking its cue from

[Chapters 3](#)

and

[4](#)

[i.e. embarking from the conclusion that the neoclassical approach to bargaining, in particular, and to rational behaviour in the context of finite dynamic interactions, in general, leads to a vicious indeterminacy that can only be suppressed by logically illicit (often hidden) moves (that correspond to the neoclassicists' *third meta-axiom*)], it asks: *Granted that the neoclassical method takes us nowhere, is there an analytical alternative?*

This is an important question. For the neoclassicist invariably defends her method by arguing, just as neoliberals tend to do in the realm of political economics, that There Is No Alternative. That, if we want a mathematical analysis of conflict, bargaining, agreements, contested contracts etc., the neoclassical model is the only option. Unless

we can show that there are analytical alternatives to the neoclassical manner of modelling these important economic processes, the neoclassical mind feels utterly unthreatened by our criticisms, regardless of how apt and piercing they might be.

This chapter has shown that it is not only possible to model analytically these phenomena in a manner that escapes the neoclassicists' logical incoherence but, additionally, that the resulting analysis is richer, more sophisticated and ultimately more ... analytical. Yet, and this is crucial to the book's theme, the type of analysis demonstrated in this chapter, despite its logical superiority to anything the neoclassical toolbox has on offer, was never seriously taken up by the economics profession. Graduate students have continued to work on equilibrium models, to assume away out-of-equilibrium beliefs, to ban patterned bluffs and, generally, to pursue a research agenda that is constitutionally incapable of embracing the creative manner in which rational people subvert the neoclassical theory of how they 'ought' to be behaving. The question is: Why?

Consistent with the rest of this book, my answer is simple: *Neoclassical economics draws its immense discursive power from the pretence that it can defeat indeterminacy*; that it can 'close' its models and pinpoint precise outcomes given the agents' initial beliefs and preferences. Any graduate student who invests time and energy on the type of model presented in this chapter will sooner or later have to confront journal editors who will ask them: 'Granted that your model offers a rich analysis of the phenomenon at hand, can you narrow down the outcome, or solution, and make it depend entirely on the model's "primitive data"?'

Any brave soul that responds 'no, not without violating the rules of logic' will be told that her paper is rejected and, consequently, face a life of drudgery as a teaching instructor in non-research tertiary institutions. Unsurprisingly, the economists that make it on the faculty boards of the good, research-based, Economics departments are the ones that bite their tongue and 'close' their models. Is it any wonder that the rich analytical framework presented in this chapter is so thin on the ground?

Appendix 5.1: Proof of Proposition 5.2

,

Section 5.2

Beginning with the *Assumption* preceding

Proposition 5.2

, the proof follows five steps:

Step 1: Negotiation Let d^F and d^U be the bargaining dispositions which base their actions on the least amount of information (that is, the smallest samples from existing records). Assume that their sample sizes are less than or equal to half the existing (H) records. Let one of the two dispositions (of F and U respectively) be slightly more keen to count in previous negotiations which ended before round t : say, $m = m_F \geq m_U$ [of course this inequality could have been reversed; in that case substitute m_F with m_U in what follows]. Note that during negotiation i (round t) the history of bargaining between F and U process is given by $\{(x_t, y_t)_{i-m}, \dots, (x_t, y_t)_I\}$

Step 2: Negotiations $i + 1$ to $i + m$

- $Prob(d^F \text{ and } d^U \text{ bargaining dispositions will be selected every time}) > 0$
- $Prob(d^F \text{ and } d^U \text{ will draw the same samples from history } H \text{ every time}) > 0$

Letting (x, y) be the best replies to these particular samples, it follows that: if φ is a potential history of exactly (x, y) demands in round t in all of the $(i + 1 \text{ to } i + m)$ negotiations, then the probability that φ will be observed in each of the $(i + 1 \text{ to } i + m)$ negotiations is positive.

Step 3: Negotiations $i + m + 1$ to $i + 2m$

- $\text{Prob}(\text{same } d^F \text{ and } d^U \text{ dispositions will be selected}) > 0$

If they are selected, they may sample from history ϕ above. In this case, their best replies to those observations are $(1 - y, 1 - x)$. Let ϕ' denote a potential history of $(1 - y, 1 - x)$ demands in round t of all negotiations $i + m + 1$ to $i + 2m$. We conclude that:

- $\text{Prob}(\phi' \text{ being observed in each of the } (i + m + 1, i + 2m) \text{ observations}) > 0$.

Step 4: Negotiation $i + 2m + 1$

- $\text{Prob}(\text{same } d^F \text{ and } d^U \text{ dispositions will be selected}) > 0$

If they are selected, the probability that they will draw samples from ϕ' is also positive. Also,

- $\text{Prob}(d^F \text{ will look back } 2m_F \text{ negotiations and } d^U \text{ } m_U \text{ periods}) > 0$

In that case, F 's and U 's best demands are $(1 - y, y)$.

Step 5: Negotiation $i + 2m + 2$

The history of demands ϕ has by now vanished from record (since $m = m_F < 2H$) – see the *Assumption* prior to

Proposition 5.2

. However, d^F can still gain access to records in which d^U consistently demanded y . The firm's best reply to that observation is to demand $1 - y$. In the meantime d^U has access to the more recent history in ϕ' in which the firm demanded $1 - y$. Its best reply is to demand y . In conclusion, there exists a positive probability that their pair of best demands is given by $(1 - y, y)$. Thus,

- $\text{Prob}[\text{a history of } H \text{ negotiations with settlements } (1 - y, y)] > 0$.

Appendix 5.2: On the construction of Tables 5.2

—

5.4

All three tables were based on the assumption that π_k is constant for all k .

Table 5.2

was derived as follows: Condition (

5.13ii

) tells us that, for $d = 0.5$, non instrumentally rational player would 'cooperate' as long as there are fewer than five coins left on the table (3.66 if $d = 0.4$, 3.428 if $d = 0.3$ etc.). The best chances for cooperation correspond to $y = 1$, in which case $f(y, d) = 1/2d$. Thus the probability of a cooperative move with 5 coins on the table equals, at most, $\pi(1 - \pi)/2d$. For this move to be instrumentally rational, $\pi(1 - \pi)/2d$ must exceed d – see

condition (

5.13i

). Clearly it does not when there are five coins left for any π when $d = 0.5$ (notice that it does when $d = 0.25$).

If $m = k - 4$, then the probability of a cooperative move can reach a maximum $[m\pi(1 - \pi)/2d]$. For this quantity to exceed $1/2$ (i.e. for a cooperative move to be instrumentally rational with $k = m + 4$ coins left), $m = 5.55$. Thus the total number of coins must be a minimum of 5.55 plus 4, which equals 10 after rounding. Similarly for the rest of

Table 5.2

;

Table 5.3

was compiled in a similar way. Finally,

Table 5.4

generalises by allowing for non-linear alignment between the players beliefs using

a probit specification for $f(y, d)$. The range of the minimum number of coins reported corresponds to a choice of the probit's two parameters such that the divergence from the linear case does not exceed one standard deviation.

Notes

- 1 Sugden (1990) concurs that bargaining theory can pinpoint uniquely rational solutions to the bargaining problem only if it assumes that such solutions exist; an assumption (which he refers to as Rational Determinacy) that he deems analytically indefensible. See also Varoufakis (1991, Ch. 5) for a similar critique of subgame perfect bargaining solutions with particular reference to the conventional literature on strikes.
- 2 For example see Hayes (1984), Hart (1989), Kennan and Wilson (1989), McConnell (1989) and Mailath and Postlewaite (1990) – and recall
[Chapters 2](#)
,
3 and
4 above.
- 3 For example, unless bargainers are prevented from exchanging offers or demands at will [e.g. if there is a minimum delay between offers as in Rubinstein, 1985, or one has the capacity to shut down channels of communication after issuing a demand as in Admati and Perry (1987)], optimal strike duration tends to zero. Then irrationality is the only explanation of why strikes occur.
- 4 The surplus over which F and U bargain equals the firm's total revenue minus non-labour costs.
- 5 For an analysis of shock build-up see Fudenberg and Harris (1992).
- 6 The literature on evolutionary stability is large and diverse. For our purposes here, a good start are Foster and Young (1990) and Kandori *et al.* (1993). For an alternative stability concept, consult Matsui (1992).
- 7 For an example of how strike dynamics can affect the relationship between a trades union's leaders and rank and file, recall
[Chapter 2](#)
.
- 8 A whole chapter will be dedicated later to this interdependence between belief and utility – see
[Chapter 8](#)
.
- 9 Perhaps surprisingly, evolutionary game theory can show that the 'fittest' do not always survive. Thus to demonstrate that some convention is more beneficial for everyone concerned, is not necessarily to show that evolution will favour it. See Dekel and Scotchmer (1992).
- 10 See McKelvey and Palfrey (1992).
- 11 See Binmore (1987), Pettit and Sugden (1989) and
[Chapter 4](#)
.
- 12 See Kreps *et al.* (1982).

6 Marxists and the sirens' song

When disgruntled Marxists reach for neoclassical economics' toolbox they get more than they bargained for

6.1 Prologue

6.1.1 Background briefing

If I had to give a single example of the exceptional discursive power that neoclassical method attained via its *dance of the meta-axioms* (see

[Chapter 1](#)

), it would be the emergence, sometime in the late 1970s and the 1980s, of a group of Marxist scholars calling themselves 'Rational Choice Marxists' (RCMs), with John Roemer and Jon Elster as their *avant-garde*. As I shall be arguing below, their proclaimed Rational Choice Marxism was no more than an attempt to cloak a Marxist narrative in neoclassical clothes.

The reason that, in my estimation, RCMs are an excellent example of neoclassicism's triumph is simple: It is one thing for neoclassical method's lure to rope in neoliberals who have no qualms with the notion that capitalism is the realm of pure exchanges between free agents. But it is an achievement of a higher order to ensnare scholars trained to think in the Marxist tradition which, supposedly, prepared them to beware static models and to scorn portrayals of capitalism as the canvas on which pure exchanges between equally free and rational agents paint their free and efficient society.

Crucial to the attraction that landed RCMs in the trap of neoclassical economics was game theory. Game theory came to the fore in the early 1970s with claims that many social theorists (often lacking the mathematical skills to investigate these claims thoroughly) found impressive. In particular, game theorists claimed to have developed mathematical models that could, at last, model phenomena that had previously defeated the efforts of neoclassical economists: conflict, strategic behaviour, oligopoly, bargaining etc.

- The crispness of the prisoner's dilemma, in demonstrating how rational people can intentionally behave in a manner that is collectively and personally self-defeating.
- The brilliance of John Nash's equilibrium concept, that cut through a plethora of potential beliefs to pinpoint the beliefs that agents could entertain without presuming that their opponents would *ex post* regret their *ex ante* expectations.
- Game theory's capacity to analyse repeated games to demonstrate, for example, how monopoly power could be maintained by the smart application of seemingly irrational behaviour on the monopolist's part.

All these game theoretical offerings proved particularly alluring for RCMs desperate to paint a fresher, more commonsensical, picture of capitalism as an inefficient system.

Of course it was not just the enticement of game theory's wares that did the trick. There was also, crucially, a great deal of (understandable) fatigue amongst Marxist scholars caused by the obscurantist narratives of the Marxist tradition which, for decades, was recanting the same old tired functionalist arguments which, to the young scholars that later formed the RCM group, sounded increasingly tedious. Put simply, they had had enough of explanations that went like this: 'Liberal democracy prevailed in the West because it was functional to the interests of capital accumulation in advanced capitalist countries'. Jon Elster (1982) explains nicely how analytical leftwing thinkers like himself would no longer tolerate such brittle explanations of phenomena on the basis of their function alone. They thought that, while functionalism was useful in, say, biology (e.g. explaining the structure of the human stomach on the basis of its function within the human body), social science required intentional explanations of the

phenomena under study. And as neoclassicism trades on intentional explanations, RCMs were naturally inclined to train their antennae toward it.

Once neoclassical theory managed to move beyond parametric choices and embrace strategic behaviour (i.e. once neoclassical economics had spawned game theory), scholars like Elster found it irresistible. Suddenly they could explain in analytical terms, based on an intentional model of agents, why capitalists often make choices (e.g. lower the workers' wages) that, at the macro level, prove detrimental to capitalism. 'At long last', RCMs thought, 'we can be scientific Marxists who no longer have to rely on obscurantist references to some 'dialectical' process and, instead, support our Marxism by means of mathematised analytical reasoning of the sort that mainstream economics was based on.'

Back then, in the 1970s and 1980s, game theory was fairly young and imperfectly understood. RCMs did not recognise that by adopting game theory they were unwittingly purchasing into the neoclassical method. For them, neoclassical economics was all about parametric optimisation (i.e. agents taking the rest of the world as fixed, before making their optimising choices), which meant perfect competition and the complete absence of oligopoly, conflict, market power etc. So, when they heard that game theory permitted the modelling of all these dynamic phenomena, they presumed that it was a non-neoclassical, scientific method that Marxists ought to press into the good and proper service of exposing capitalism's deep-rooted ills.

Alas, in adopting game theory, RCMs were espousing the highest, most rabid form of neoclassicism (as the previous chapters have demonstrated). The price they paid for the neat 'solutions' game theory offered them, to issues that as Marxists they had traditionally grappled with, seemed worth paying. Only it was not. On the one

hand, the adoption of a game theoretical, deeply neoclassical, framework of analysis caused them to forfeit the analytical advantages that Marx conferred upon them: the notion that labour markets can *never* be a realm of free and pure exchanges; the idea that human rationality and freedom are radically indeterminate. On the other hand, they ended up with a series of game-based models whose outcome was only allegedly determinate; but whose determinacy crumbled once a critical light (like that shone by the preceding chapters) fell upon them.

In summary, in order to gain determinate mathematical models that confirmed their Marxist conclusions regarding capitalism (e.g. that capitalism is constitutionally wasteful of human capacities), RCMs forfeited the indeterminacy of the human agent; the meaning of her rationality and freedom in particular. Tragically, this hefty price was paid for nothing; for as we saw in previous chapters, and will see again in this one, the determinacy neoclassicism lured RCMs with depended entirely on the *third meta-axiom*; i.e. on logically impossible axioms introduced underhandedly in order to 'close' the models and provide, with no rational explanation, the promised determinacy. In the end, RCMs were left with neoclassical models that, first, were as indeterminate as any and, secondly, had been rendered (courtesy of the RCM's acceptance of neoclassicism's three meta-axioms) utterly disconnected from really-existing capitalism.

6.1.2 Sketch of the chapter

Rationality and freedom are 'strange' concepts. Just like beauty, you know *when* you are in their presence but it is impossible precisely and analytically to define them. Indeed, any attempt to define them diminishes them brutally. While as real as anything that is crucial to the human condition, freedom and rationality are deeply indeterminate and resistant to open-and-shut definitions.

Neoclassical economists are, of course, not the kind of thinkers that would lose sleep over the loss of 'meaning' that follows precise definitions of notions such as beauty, truth, rationality or liberty. If the success of their discourse depends on defining rationality fully (e.g. in order to identify utility maximisation with rationality), define it they

will. Similarly with freedom. And if what we are left with are thin-as-a-needle notions (of freedom and rationality), so be it!

Well, the result of the neoclassical penchant for defining fully Reason and Liberty are (what I have been referring to throughout this book as) 'instrumental rationality' and 'negative freedom'. Instrumental rationality is simply the assumption that a person is rational to the extent that she deploys her available means efficiently in order to achieve her pre-existing, exogenous, objectives. And when these objectives can be captured by means of a well-defined utility function, then instrumental rationality boils down to a capacity for utility maximisation (given one's exogenous constraints). As for 'negative freedom', it is nothing more than the absence of constraints, which translates into the normative notion that agreements are free (and, thus, fair) if the parties to it have consented to them.

Section 6.2

argues that both these definitions are crude and cause those who accept them to lose sight of what it truly means to be a rational and free human

being; including Marxists (the RCMs) who espoused such definitions in order to gain access to the game theoretical toolbox of neoclassical economics.

Section 6.3

presents the type of analysis that results when one adopts game theory's method in order to demonstrate that Karl Marx had a point in thinking of capitalism as an inefficient mode of production, distribution and exchange. It argues that, perhaps inadvertently, the result is a strange 'bird' of a theory; one that is best described as neoclassical Marxism. And since all neoclassicism, as

Chapter 1

explained and

Chapters 2

—

5

illustrated, ends up in a mire of radical indeterminacy, the same applies to RCM or neoclassical Marxism. Indeed, the only route available to RCMs for 'tying down' their models, and producing the determinate stories they were after, was to perform one of the moves in the *dance of the meta-axioms* (e.g. the backslide in the figure of

Chapter 1

) that purchases determinacy at the exorbitant cost of logical incoherence.

Section 6.4

proposes a different course. It suggests that indeterminacy must be embraced in the manner outlined, even before Marx, by Hegel. Reason is then to be conceptualised as a *process* of rationality-creation which codetermines the individual's motivation and the rules of rational choice. It is a process that is characterised as much by logic as it is by experiments with alternative rationalities.

¹

And it is a process that neoclassical theory, loyal to its static spatial metaphors, cannot keep up with. Then

Section 6.5

turns to another set of sirens that have proven quite alluring in the past few decades: that of Postmodernity, which castigates anyone who may be trying to theorise rationality and freedom. The postmodern view is that words such as 'Reason' and 'Liberty' are signifiers that signify nothing more than transient figments of our imperfect language. In short, like all metanarratives, they are indeterminate concepts best treated as such, leading us to the conclusion that nothing concrete can be said about them. My response to this depressing claim is that, while indeterminacy must be embraced, we do not have to lose the hope that concrete theory concerning our social world is possible.

Section 6.6

presents a suggestion of what such a concrete theory, which does not eschew indeterminacy, might look and feel like. Central to my suggestion is an encouragement to entertain the possibility of indeterminacies that only history can resolve. Note that this does not constitute a slide into postmodernity. The latter canvasses the non-availability of deeper truths, the uselessness of large scale explanatory systems, and the view that ideas are interpretable only in terms of their past and present cultural relevance. At every point in time, the dialectical Reason proposed in

Section 6.6

is capable of grasping the logic implicit in those episodes of thought which make up its own pre-history. What it does not do is to specify *ex ante* the exact path on which it will tread. The dialectic's horizon is a state of perpetually open possibility that human praxes determine in real time. (Or so I claim!)

6.2 Freedom within Reason

Does being free mean that we are not unfree and are we rational if we are not irrational? If liberty and rationality are notions not dissimilar to those that nature throws out, then their definition is possible by means of negation provided they and their opposites are mutually exclusive. In the same way that a substance is organic if it is not inorganic, a woman will be thus declared free if she is not unfree or a deed rational if it is not irrational.

An initial criticism of this definition may throw the spotlight on its absolutism. A person can find her environment to be more or less oppressive, or act in a manner that displays elements of irrationality without being downright stupid. Indeed, even neoclassical economists have argued that the problem with their discipline is that it has trouble recognising degrees of irrationality or unfreedom. In the case of the latter, economic models have been castigated for their failure to capture the loss of autonomy due to unequal wealth or property rights, and in the case of the former they have been criticised (as the previous chapters will testify) for making unrealistic assumptions concerning the ability of agents to think clearly.

To illustrate the above and motivate the forthcoming point about Liberty and Rationality on which this chapter turns, suppose we have a field encircled by a fence. Inside we have freedom or rationality and outside we have their opposites aching to get in but prevented from doing so by the fence. Neoclassical economic theory in particular and liberalism in general are completely taken by this essentially Humean metaphor. The tentative criticism identified in the paragraph above suggests that the demarcation of freedom and rationality from tyranny and senselessness may not be so neat. Parts of the fence have caved in and there is a grey area in which the two concepts live in an uneasy symbiosis with their opposites.

The mixing is not complete, as there are inner defences that do not allow the barbarous outsiders a complete walk-over, but it is serious enough to warrant studies of bounded rationality (in the context of limited computing ability) and of degrees of liberty (in terms of distributive justice). No doubt these amendments accommodate the initial criticism by conceding that some tension between rationality and liberty on the one hand, and irrationality and unfreedom on the other, must be entertained. From the outset, social theories are built upon the assumption that the fence is intact and then, once the social world is better understood, the assumption is relaxed and new insights are sought as the fence begins to baulk. Nevertheless, the spatial paradigm is at the centre of all 'established' theory.

The criticism advanced here goes deeper as it rejects the very possibility of properly understanding rationality and freedom in terms of geographical metaphors. By the very nature of these metaphors (e.g. the fence) their portrayal of liberty and rationality implies that the social and historical milieu of the persons who will be endowed with these gifts is independent of such notions. It is customary, thus, to take no account of

the simple proposition that rationality and freedom demand not only a physical capacity to act freely and rationally but also that the agent has reached a certain level of social development and is *conscious* of these notions. And here is the rub: Before I can do something with my freedom of speech I must have something to say. If my faculties permit me to attain my objective, I must have an objective before my action is deemed rational. Moreover, I must be wise in the way I choose my objectives; a thought that the neoclassical mindset greets with stunned incredulity.

Rooted as neoclassical economists, and their handmaidens in the rest of the social sciences are in a naturalistic perception, of which the fenced field is one example, they are far less demanding of the rational agent. To coin another naturalistic metaphor, which lies behind the neoclassicist's perspective on freedom, the main condition for a satellite to break loose from a planet's gravity is that its vectorial speed exceeds a certain threshold. Either its speed exceeds the threshold or it does not. Though we may say that the satellite has been set 'free' if it does, it is ludicrous to mistake the metaphorical resemblance between the satellite's 'freedom' for the freedom of human agents for something more profound. It is equally absurd to succumb, as neoclassicists do, to the temptation of identifying the efficiency with which targets are reached with rationality. The former is an adequate rule to use in ballistics but quite inappropriate as an inclusive guide to rational behaviour.

Unfortunately, the metaphorical definitions of Reason and Liberty that we observe as part of the liberal and neoclassical narratives *seem* correct because our language permits associations between notions such as 'free fall' and 'free speech'. The danger comes from our tendency to accept analogous definitions for concepts whose analogy springs from the common metaphor our minds utilise in order to attain comprehension. If the analogy is epiphenomenal, as it certainly is in these two cases, it is bound to cause serious confusion. For instance, the conditions that must hold before the phenomenon of an object travelling through the ether is definable as a 'free-fall', can be described without reference to the object itself. In other words, 'free-fall' is definable a priori and by means of a natural science rule that is independent of the object. In contradistinction, the phenomenon of 'free-speech' is not. Any attempt to construe it without reference to the determinants of what a person has to say, is pregnant with the danger of describing an instance where a voice synthesiser recites a speech randomly selected from its memory banks as a manifestation of free speech.

Similarly, although we can specify a priori measures of the speed with which a computer performs a calculation, it is impossible to use similar criteria for assessing the rationality of humans, unless we are happy to think of rationality solely in terms of the axioms that usually reside in the first few chapters of a neoclassical microeconomics textbook (I refer to this, throughout this book, as 'instrumental rationality'). It transpires that the type of definition of liberty and rationality espoused reveals a social theory's make-up. To define these notions a priori, and by means of metaphors that exclude the social and historical background, is to dehumanise and impoverish social theory.

In the following pages I focus on the implications of this dehumanisation of social theory. By depleting the meaning of rationality and freedom, bourgeois thought achieved two things: First, it imposed its own perception of the two notions on all people and at all times and, secondly, it paved the way to a celebration of their loss. The former obtained as axiomatic definitions were put into place which, although philosophically weak, turned the bourgeois urge to accumulate into the major determinant of what it means to be rational and free. The latter resulted from the discontent caused by the repercussions of the axiomatic

approach. By denying their substantive meaning, the choice of metaphorical narrative for the two notions (freedom and rationality) strengthened the hand of those who wish to claim that the loss of abstract theories of Reason and Liberty is a blessing

in disguise. Should bourgeois thinkers worry or will they take this postmodern twist in their stride? On the one hand, they will feel threatened because their positivist models will no longer be presentable as positive analyses of the social world. On the other hand, the postmodern reaction to Reason and Liberty is functional to bourgeois thought if the postmodern revelation that the latter has no clothes is a better disguise of its nudity.

By the chapter's end I hope to have shown that, if we are to reclaim freedom and rationality, we need to distance ourselves from the axiomatic definitions which are based on ontologically static binary oppositions between notions and their opposites. Moreover, we must also resist the sirens of Rational Choice (or Analytical, or Neoclassical) Marxism, as well as those of Postmodernity, and turn to old-fashioned dialectical materialism. However, our arrival at dialectics must complete a journey of renewal and not just a journey home. Neoclassical theory may be incapable of enlightening the *meaning* of rational or free choices, but it does illuminate the way in which bourgeois thought generates insurmountable internal contradictions. Postmodernity may fail to expose the folly of bourgeois naturalism without debunking progressive thinking, but it does point to a tendency toward metaphysical determinism by those of us (and here I mean Marxists of all different sorts) who have been sloppy with their dialectical reasoning. There is a lot to learn from the ability of bourgeois theory to undermine itself in order to frustrate the development of a progressive social science.

6.3 Neoclassical Marxism

If one is to define rationality before anything is said about the human subjects who will then be endowed with it, one is forced to adopt instrumental (means-ends) rationality. Precisely as neoclassical economists do. For, if a more substantive type of rationality were to be admitted, one would have to know the societal values surrounding agents (e.g. Kant) and their history (e.g. Hegel) before discussing their Reason.

²

Nevertheless it is often said that although instrumental rationality is insufficient (in that it needs to be supplemented by other forms of Reason), it can still offer some useful local explanations when the preferences of agents, as well as their environment, are stable. Hargreaves-Heap (1989a), for instance, argues the case for retaining instrumental rationality without relying on it exclusively. On a similar note, Carling (1990) insists that local explanations with given objectives and social structures are not to be scorned. Indeed, he sees in these explanations the foundation of Rational Choice Marxism and invokes them as ammunition against those who argue that any use of instrumental rationality for the purposes of historical explanation is illegitimate – e.g. Ellen Maaskins Wood (1989). Alas, what he calls Rational Choice Marxism can be shown to be a variant of neoclassicism in which the maximising agents are labelled 'workers' or 'capitalists', as opposed to just Jack, Jill, the firm or the trades union.

Such defences of the use of instrumental rationality for 'subversive purposes' traditionally cite particular social situations where persons find themselves in circumstances describable by a prisoner's dilemma that offers an appealing explanation of why instrumentally rational agents may act in a self-defeating manner. For example, one can argue that *it may be more rational to be less instrumentally rational* (an essentially Kantian proposition) or, in a revolutionary flourish, go further and develop an argument in favour of *collective action as an instrument for changing the objectives of individuals* and thus transforming their social relations in ways conducive to superior social outcomes. Unfortunately such transitions cannot be analysed in terms of instrumental rationality since the latter needs unchanging aspirations and reasoning before it is operational. Granted that the offered narrative is attractive, is it more than a restatement of what Hobbes, Hume, Rousseau and Marx have already debated in splendid prose? I think not. But what of the argument that it may be a useful

restatement of what we already know? I would be happier to accept the validity of this statement if it were not for the serious danger that the neoclassical method these 'subversive' narratives come wrapped in poses serious threats to the logical coherence of the offered analysis.

The foremost advocates of asocial and ahistoric a priori definitions are, not surprisingly, neoclassical economists. When asked to produce evidence that genuine and original social explanation can be procured as a consequence of models in this manner, they quickly present us with the notion of the *Nash equilibrium*. For the uninitiated, suppose you are involved in a strategic situation in which you are compelled to make a choice. However, the outcome of your choice depends not only on what you choose but also on what others choose. In cases where none of your choices is dominant (that is, is best regardless of what others do), what you ought to choose critically depends on what you think that the others think that you think ... *ad infinitum*. The Nash equilibrium is what game theory has to offer as a way out of the infinite regress.

Interestingly, this game theoretical notion has attracted attention from theorists whose agenda is quite different. The so-called RCMs have argued that the analytical power of the Nash equilibrium should be harnessed by progressive social scientists in order to dissect social relations and establish a micro-foundation for historical studies of social change. According to Jon Elster, for example, the Nash equilibrium offers a way of understanding complex social interaction with a simplicity reminiscent of the irreverence with which Alexander the Great approached the Gordian knot. By contrast, Ellen Meiskins Wood (1989, 1995) is concerned that RCMs assume the structures that need to be explained and that, by the time the Nash equilibrium is called upon to deliberate as to which is the best choice of rational actor, there is very little left to explain.

Unlikely as it may seem, there is an implicit consensus between neoclassical economists, RCMs and critics such as Wood who has a distaste for attempts to illuminate history by focusing on the decisions of individual or collective agents in a neoclassical framework. The consensus concerns the capacity of the Nash equilibrium to explain how agents will behave given their objectives and environment. Wood may disagree with the RCMs about the value of developing such

explanations if the price that has to be paid is a neglect of classical Marxism, but she does not challenge the proposition that, given objectives and environment, history-free a priori definitions of rationality can yield such explanations.

Similarly, RCMs are prepared to accept wholeheartedly the usefulness of tools like the Nash equilibrium even if they wish to put them into the service of subversive social science – so-called heterodox (i.e. non-neoclassical) economics.

In this section I want to agree with Wood but also to take her repudiation of the neoclassical paradigm further by challenging the above consensus. For it is not only that neoclassical theory assumes most of what it tries to explain (as Wood correctly remarks) but that, even after it has made its assumptions, its explanatory power is severely overstated. The cause for this failure is the impossibility of a sensible ahistorical a priori definition of rationality and its main repercussion is that we can no longer confidently expect to know what a rational agent will do even if we know her objectives and the precise environment in which she functions. In short, and in resonance to the theme of this book, the neoclassical model of men and women, even when it features some unique Nash equilibrium, throws up irrepressible indeterminacy: it fails analytically to pin down the outcome of the interactions it places under its microscope.

The best way of illustrating the argument is by means of a simple example which we have also seen in the context of

Chapter 3

(see

Section 3.3

). Neoclassical game theorists recognise that not all social interactions (in their language, games) have clear-cut solutions. If there exists more than one equilibrium strategy (or action), it is unclear what a rational actor will do. Indeed, a burgeoning literature is trading on the multiplicity of equilibria and the resulting need for selecting between them. RCMs have been made aware of this difficulty and some have constructed elegant analyses of collective action and evolution based on the non-uniqueness of equilibrium solutions.

However, this recognition is founded on the belief that, in social interactions where there exists a unique equilibrium, the outcome is an open and shut case. It is a belief that, alas, cannot be sustained simply by assuming instrumental rationality. Thus, it is a belief that must be deflated.

Consider a heuristic game (of the sort employed by RCM) where the working class is pitted against capitalists and each side has three strategies at their disposal. Strategies range from the cooperative (within capitalist confines) to the combative. The workers' second strategy is to struggle for a higher portion of the surplus without attempting to overthrow capitalism, while their third strategy is openly to contest the bourgeois state and its property relations, i.e. Lenin's revolutionary option. Capitalists on the other hand must choose between retaining a liberal-democratic environment, enforcing anti-labour legislation which bans strikes and, lastly, calling in the military, thus discarding the liberal-democratic cloak of legitimacy altogether.

Before the reader despairs at the above description, let me confirm that it is meant as an entertaining rather than a historically useful example. Nevertheless, it serves its role of revealing the problem with the RCMs' attempt to deploy neoclassical methods for the purposes of radical politics. Similarly, game theorists will not mind my choice of labels for the various strategies below since they claim to give answers based entirely on payoffs and timeless a priori rationality, and utterly independently of the interaction's social and historical context. As for the RCMs, they too can ill-afford to challenge my narrative in view of their adoption of the analytical tools of game theorists. As for the rest of you, dear readers, I plead that you bear with me for a little longer.

Table 6.1

A class war game

	<i>Capitalists</i>		
	<i>1. Liberal democracy</i>	<i>2. Repression (e.g. ban strikes)</i>	<i>3. Military repression</i>
Workers			
1. Cooperate	10, 100 ⁻	5, 90	+5, 80
2. Struggle for higher portion of surplus	50, 50	+15, 80 ⁻	2, 70
3. Revolution	+150, -50	10, -10	-10, 0 ⁻

In

Table 6.1

we have a typical game where no strategy dominates – that is, each strategy is rationally playable depending on the agent's belief. For instance, if workers believe that capitalists will choose strategy 1, then their best response is 'revolution' (strategy 3). If they anticipate that capitalists will introduce antiunion legislation (strategy 2), their best action is to intensify the struggle within the capitalist framework (strategy 2). Note that I mark the best responses of the row player with a plus sign and those of the column

player with a minus sign. Quite clearly, workers will act according to what they think that capitalists think that workers will expect capitalists to....

This is a good example of the type of analysis to be found in rational choice and game theory. The starting point is to assume that payoffs can potentially capture the motivation of players (see Hollis, 1990 for an objection that I shall return to later). If the choices of players are to be rational, they must be based on expectations rather than played thoughtlessly. The problem is how to choose the right expectations. Neoclassical economists, following John Nash, observe that there exists only one outcome which results from choices which confirm the expectations that support them: outcome (15, 80) where both sides have chosen their second strategy. Workers would only play **2** if they expected capitalists to play **2** and vice versa (observe the coincidence of the plus and minus signs at that outcome). It so happens that at (15, 80) both of these expectations are confirmed. Moreover, (15, 80) is the only outcome that confirms both players' expectations – that is, it is a Nash equilibrium.

It is easy to spot that (15, 80) is the unique equilibrium outcome. *Any* other outcome can only be reached if at least one of the two sides is motivated by expectations that are bound to be proved wrong. For example take outcome (150, – 50) – that is, the case where there is a revolution while the capitalist class tries to retain its dominance by sticking to liberal-democratic institutions

and processes. For the workers to rebel they must have expected capitalists to play their first (liberal) strategy. At outcome (150, –50) their expectation has been proved correct. However, capitalists will only play that strategy if they expect workers to cooperate. Thus, at outcome (150, –50) capitalists will have regretted their decision as it was based on an expectation that was disproved by the facts.

'So what?', one may rightly ask. What if (15, 80) qualifies as the Nash equilibrium by being the only solution which confirms the expectations of both agents? This observation is of analytical value only if we believe that rational agents gravitate towards choices that invariably make them feel vindicated vis-à-vis their expectations *after history has unfolded*. Notice that no evolutive argument in favour of a convergence between beliefs and actions is possible here in view of the absence of historical time (the game is only played once). John von-Neumann and Oskar Morgenstern (1944), the founders of game theory, wrote:

We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable. But there is ample evidence from other branches of science that it is futile to try to build one as long as the static theory is not thoroughly understood.

The influence on these pioneers of natural science is evident. Unfortunately it is an influence that leads the analysis astray since rationality (i.e. the motivating force behind human actions), unlike natural forces, is cognisable only within a social complexity. It is the argument of this paper that attempts to build a sound static theory of rational choices is next-to-futile. Moreover, the dynamic theory built upon any such static theory is bound to inherit the problems of its foundations.

If we believe, as I think we should, that social interaction can lead to instances where rational agents regret decisions they have made (something even chess masters often do), then the discovery of a unique Nash equilibrium sheds no light on the question of what constitutes a rational choice.

5

And yet neoclassical theory axiomatically imposes the condition that rational agents *must* have expectations that are *always* confirmed by history. Based on this assumption, it goes on to build the magnificent castles that are to be found in the prolific game theoretical literature. Unfortunately, they are built on sand as there is no reason why rationality ought to engender consistent alignment of the agents' beliefs. Let us make no mistake here: The only way one may take (15, 80) to be the unique solution is

if one assumes that each and every agent chooses *ex ante* in a way that their expectations are confirmed *ex post*. Put bluntly, neoclassical theory can home in an outcome as *the* solution if and only if it assumes away the most important aspect of strategic interdependence, namely the uncertainty in the mind of players about whether their conjectures are good ones or not. As Hollis (1990) puts it, if our agents were computers ‘... we would be asking whether they were two computers or one with interconnected routines’.

To give a feel for the alternative outcomes that may eventuate, consider the following train of thought that will lead workers rationally to initiate a ‘revolution’:

We shall rebel (strategy **3**) because we expect capitalists to choose their ‘liberal-democratic’ strategy (strategy **1**). If they knew that we contemplated ‘revolution’, they would of course choose to suppress it (Strategy **3**). However, we think that they expect us to be fearful of this possibility and to ‘cooperate’ for this reason. Hence, we think that they do not anticipate a ‘revolution’, as they are confident that we dread the prospect of a crushed revolution [that is, outcome (–10, 0)]. They will therefore, we believe, choose the liberal-democratic road. Hence our best course of action is to rebel.

Will their expectation that capitalists are about to choose the ‘liberal-democratic’ strategy be confirmed? Perhaps, but not necessarily. The condition for this to occur is that capitalists think as follows:

We pursue the ‘liberal-democratic’ road because we anticipate that the working class will ‘cooperate’. The reason why we expect this is because we believe that they fear that we are about to unleash a military coup (strategy **3**) expecting that they will rebel. And why do they think that we expect them to rebel? Perhaps because they think that we fear that they believe that we will play ‘liberal’ in which case we should expect them to rebel!

If the above capture the two sides’ thoughts, then outcome (150, –50) will appear as the result of perfectly rational choices. That one of the two sides (in this case the capitalists) will eventually realise that its conjectures were false, is a normal byproduct of the inescapable indeterminacy occasioned by social conflict. The important point here is that, since there are different consistent trains of thought which support each and every outcome in this game (including the equilibrium), an a priori definition of rationality cannot by itself elucidate this game. Even though the structure of payoffs (that is, the social context according to Wood) is given and there is a unique equilibrium, there is no plausible theory of what will happen. [In

[Appendix 6.1](#)

I list the rational trains of thought that may lead either side to choose any of its three strategies.]

In conclusion, the neoclassical theory of rational choice is not at all about discovering the rational *core* of human actions and thoughts. It can narrow outcomes down but only if it imposes the condition of belief alignment without regard for the canons of rational discourse. But then it should be referred to as Telepathic Choice Theory and RCMs, if they remain committed to the Nash equilibrium, should become Telepathic Choice Marxists!

6

6.4 From instrumental rationality and axiomatic liberty to Marxian praxis

Fencing irrationality out in genuinely strategic situations does not work. We saw this in [Chapters 3](#)

and

[4](#)

, where the case was made that the genuinely rational agent knows how profitably to undermine the dividing line between rational and irrational actions. To define, and

delineate, the 'fence' that divides the rational from the irrational acts requires that the theorist augments the a priori assumption that players are rational with a logically illegitimate assumption regarding their conjectures. Nor can we define freedom as the moral space which is kept clear of unfreedom. Those who have tried to define liberty in negative naturalistic terms (e.g. Robert Nozick, 1974) either feel the need to transcend their own definition at some level (see Isaiah Berlin, 1953) or end up with what Robert Paul Wolff (1980) refers to as a conception of rights and liberty which would '...immobilise us all, making us much like a bizarre gathering of morally musclebound rights freaks, lovely to look at, but unable to lift a finger for fear of encroaching on one another's moral space'.

Freedom as a concept is thus brought to us by a spatial metaphor whose roots can be traced to the eagerness of landlords to keep trespassers at bay. In this sense it is a metaphor that lends itself to the pure exchange paradigm to be found at the heart of liberal contractarianism. To be free is thus to make rational choices unimpeded and to trade at the market place at will – the bourgeois ethic suitably distilled into a conception of liberty. Hegel describes the freedom emanating from market exchanges better than recent advocates of economic liberalism (e.g. Milton Friedman, 1962) when he writes: Only because the other sells his good that I also do so; and this equality in the thing as its interior is its value, which has my complete consent and the opinion of the other – the positive mine as well as his, the unity of my will and his.

[Hegel (1942)]

7

Before investigating further the centrality of market relations to the notion of freedom, it is interesting to recall what makes the ahistorical a priori definitions of liberty and rationality so attractive to liberal theorists. In Hobbes, freedom and rationality are the two prerequisites for survival. Reason helps one find the best response against the barbarians living outside one's person, and freedom ensures that one will not have constantly to look over one's shoulder. In Locke, the spatial metaphor manifests itself vividly in his famous *proviso*

8

where the allocation of property rights over unclaimed resources is determined. In both cases, the primary objective is to determine moral laws which leave as little room as possible for others to meddle with one's person. Market exchanges fall naturally into place as they are seen as voluntary and conducive to the maximum degree of socialisation while permitting only the minimum interference with one's 'space'. Human identity is thought to pre-exist the socialisation phase and, consequently, freedom pre-dates the entry of social agents on stage. Thus the proclivity toward a priori definitions

of rationality and liberty that are not informed by history or by the social linkages between the agents.

Of course, the reason why this line of argumentation is problematic is that identity and personhood are a product of, rather than a prerequisite for, socialisation. If we are to erect the fences of Reason and liberty before the agents appear, then however well shielded they may be from each other they will be compatible with neither rationality nor freedom. Hegel may have celebrated the market relation as a liberating institution but he justifies it in terms of its capacity to forge a theoretical and practical relation between self and other. In contradistinction to Hobbes, Hegel points out that survival is a prerequisite for those other things that constitute social identity.

Self-consciousness attains satisfaction in another consciousness. [Hegel (1931)]

His perception of the market relation, although utterly sympathetic to that concept, turns on the thought that pure exchange between economic actors allows them to reflect upon each other and thus to become who they are.

The concrete return of me into me in the externality is that I, the infinite self-relation, am

as a person the repulsion of me from myself and have the existence of my personality in the being of other persons in my relation to them and my recognition of them which is mutual.

[Hegel (1942)]

The market becomes the arena in which freedom prevails, not because it ensures non-intervention by one on the other, but because it sets the scene for the dialectic of recognition between agents. Before a social encounter takes place, an imposition of naturalistic conditions for rights, freedom and rationality turns such concepts into impediments to human subjectivity. It is the predicament of liberalism that, in its efforts to offer an ahistorical definition of the ultimate human goods, it does away with the subjectivity that makes those goods important. If a market exchange is therefore seen as nothing more than an exchange which leaves the agents wealthier but ontologically identical as before,

9

then the contract cannot have any moral weight.

Hegel shows that the commitment of agents to honour contracts forged at the market place, develop precisely because they have rights which are not their private property and which cannot be sold freely. If freedom was definable strictly by the voluntarism of buying and selling, then every aspect of human subjectivity ought to be a commodity. In that case, no contract would have moral authority for the same reason that the slave cannot offer the master meaningful recognition as long as slavery entails complete subjugation. We must conclude that the moment we accept that human subjectivity is shaped by social interaction, freedom is not possible prior to socialisation. This insight complements the thought from the previous section that the process which shapes agents' perception of gains and of each other also shapes their normative expectations which trigger particular trains of thought and rational decisions.

Of course, one can remain a liberal without rejecting the proposition that freedom must be important for non-voluntaristic reasons and also that rational action is irreducible to instrumental procedures. Following the lead of J. S. Mill, John Rawls (1972) accepts the argument that private contracts must derive their legitimacy from somewhere other than further private contracts, and suggests that the way forward is to establish whether free and rational agents would accept the principles under which society is to be organised. In an ingenious twist whose purpose is to retain the ahistorical a priori definitions while augmenting them with public rights, he invents the 'veil of ignorance' behind which agents will decide which societal principles are legitimate and suitable as a social contract. He argues with great elegance and skill that it is *because* of asocial and ahistorical a priori freedom and rationality in the original position that the socialisation of agents, from which the legitimacy of the market obtains, materialises. Of course, the proof depends on accepting the original position in which pre-political agents can begin rationally to socialise without knowing their political roles.

The irrepressible problem here is that the moral legitimacy of pure exchange as a guarantor of freedom relies entirely on the choices made by agents before the market exchanges commence. Before the social context is collectively chosen behind the veil of ignorance, it must be demonstrated that the final choice was made after *all* potential alternatives were considered. Unfortunately, even the most imaginative of peoples cannot transcend *as a whole* their society readily and consider in a vacuum alternatives that history has not yet shaped. As Bob Sugden (1989) put it in a related debate, 'the belief that one ought to follow a convention is the product of the same process of evolution as the convention itself'. Ancient Athenians, for example, if asked to consider alternative socioeconomic organisations behind Rawls' veil of ignorance, would have probably come up with a social contract involving some form of slavery. This would not make slavery a characteristic of civil society with uninterrupted legitimacy throughout history. It took the praxes of Spartacus and countless others to

forge a sustainable perception of a slavery-free society. We must therefore conclude that the Rawlsian defence of the liberal tradition fails to redeem the liberal definition of freedom and rationality.

10

In other words, it fails to support the possibility of sensible all-inclusive axiomatic definitions embedded in logical, as opposed to historical, time.

If it is the social location and the history of agents that resolves social puzzles (like the one in

Table 6.1

), the social terrain becomes the locus of human actualisation. It is on this terrain that they gain the subjectivity and self-consciousness which makes freedom at all possible. In capitalism it is the freedom of individuals *qua* property which promotes their mutual recognition and thus development. Hegel endorses capitalist market relations not because they *respect* human freedom and allow rationality to work properly (*à la* Hobbes and Locke) but because they are the culmination of a historical process which *creates* freedom and rationality. However, for Hegel it is important that market relations treat buyers and sellers symmetrically if freedom is to be made available regardless of social position. In the market for apples and oranges there is little doubt that there is no systematic force working against such symmetry (or in Hegelian language,

recognition of self by other). It is in the market for labour that things become tricky. Marx (1972) writes:

In the market, as the owner of the commodity 'labour-power', [the worker] stood face to face with other owners of commodities, one owner against another owner. The contract by which he sold his labour-power to the capitalist proved...that he was free to dispose of himself. But when the transaction was concluded, it was discovered that he was no 'free agent', that the period of time for which he is forced to sell his labour-power is the period of time for which he is forced to sell it, that in fact the vampire will not let go 'while there remains a single muscle, a sinew or drop of blood to be exploited.'

[*Capital*, Vol. 1,

Chapter 10

]

In a society whose wealth is produced in a market in which the seller of human labour finds herself, most of the time, unable to indulge in reciprocal recognition with the buyer, Hegel's moral description of the marketplace breaks down. For how can the dialectic of recognition proceed when production is based on the non-market exchange that ensues the labour contract? Once the worker enters production, the market paradigm evaporates and the extractive power that Hegel saw markets as putting an end to, returns with a vengeance. Any society incorporating a wage system cannot, according to Hegel's own principles, be genuinely free.

C. B. McPherson (1973) revamps the liberal definition of freedom by arguing that, instead of defining freedom as the requirement that one must consent to anything that is taken from her (the invocation of the fence metaphor), we ought to define it as one's freedom not to consent to such a transaction (a thought incompatible with any spatial metaphor). Presented with a contract from a potential buyer who wants something from us, the litmus test of the purity of the exchange is whether we have the option to say 'no' and not whether we actually say 'yes'. One is free to turn down a contract provided one has alternatives. If my alternative to signing a declaration passing all of my property to the supplier of a glass of water (while on the verge of collapse in a desert) is dehydration and death, then such a contract is hardly a case of 'free exchange' *à la* Nozick or reciprocal recognition *à la* Hegel. One must have alternatives before one is free. And since what constitutes a reasonable feasible set of options is historically determined, all that Macpherson's definition has done for the moment is to restate Hegel's opposition to ahistorical axiomatic definitions. However, it does offer us the

opportunity to go further.

If capitalist social relations are marked by an incomplete market for labour,

¹¹

then one wonders why workers are prepared to enter such a market. Are they not free to choose? Macpherson's reformulation of liberal freedom emerges as relevant since they may be doing this not because of their freedom to agree but because of their unfreedom to do otherwise. Moreover, by agreeing to enter such a market, they give away the right to be part of continual market relations. From the moment they sign the contract, the rest of their working experience is not at all a market

relation between them and the capitalist. In an important way, workers freely sell their labour. In an equally important way, they are unfree to do otherwise.

Any notion of freedom that cannot handle this conceptual tension misses the point. As a social organisation, capitalism encourages workers to give up public rights which they must have if they, as well as the capitalist, are to achieve the mutual recognition which, according to Hegel, is a prerequisite for freedom in society. The meta-narrative of liberalism (and especially neoliberalism) seeks to gloss over this difficulty by explaining liberty by means of a metaphor that leaves no room for such tension. However, the moment history is introduced in the liberal discourse on freedom (e.g. Hegel) it becomes evident that the project has failed. The only reason why workers systematically give up their right continuously to exchange (that is, they enter the wage system) is that this is their only access to means of production. The private ownership of the means of production may endow capitalists with extractive power over the workers but, in Hegelian eyes, both exploiters and exploited miss out on becoming free.

6.5 Postmodernity's sirens

If the above is correct, ahistorical a priori rationality proves insufficient in strategic circumstances,

¹²

and specific socio-economic environments do not confer freedom if freedom is to be construed in terms of a spatial (fence-like) metaphor. This is the point of departure for Hegel and for Marx who see both Reason and Liberty as products of the historical process.

¹³

It is also the point of departure for postmodern writers like Michel Foucault, Jacques Derrida and Jean-Francois Lyotard who have been arguing that the reason we fail to discern the unique rational choice in interactions like that of

Table 6.1

is because of the emptiness of the signifier 'rational'. The weakness of our language, they claim, is responsible for creating the need for a demarcation between rationality and irrationality along the lines of a naturalistic paradigm.

Speaking of the fence as a metaphor (in quite a different context to ours), Derrida (1973) writes that 'no border can be guaranteed inside or out', implying that every border marks a difference which, though real, ought not to be confused with a demarcation which has the power to define what it separates. Foucault (1967) adds that we ought to abandon the search for meaning of terms that have none and recognise that there is plenty of 'Reason in Madness'.

¹⁴

As for freedom, hiding behind fences that supposedly guarantee our negatively defined liberty, is futile. It is so because the metaphor of the fence, as indeed any metaphor, is no more than a figment of our language which requires such metaphors in order to formulate concepts but is immediately hijacked by those metaphors and therefore loses its access to meaning.

Postmodernists do not spare anyone's metaphor. Having ridiculed the liberal fence, they are equally deft at deconstructing Hegel's metaphorical depiction of history and to

portray it as a discontinuous river majestically proceeding towards the sea of reciprocal recognition. They are so inclined because the postmodern critique hits at the heart of foundationalism: the method of starting from *some*

assumption about the human condition which is then taken as self evident.

15

Neoclassical game theory can be disparaged, in this sense, on the basis that it is *founded* in this manner, rather than on the basis that there are better foundations of it to have. This leads to an equally fierce attack on *any* meta-narrative, leaving history as no more than a chain whose links are different versions of the present.

Confronted by the question of exploitation that concerns Marxists, the postmodern mind comments wryly that we are all simultaneously oppressors *and* oppressed in exactly the same way that our motivated actions are at once irrational *and* rational. Postmodernity interprets the loss of the concreteness of the concepts of freedom and rationality as the inevitable decomposition of the metaphors underpinning all foundationalist social theory. Crucially, it celebrates this loss and warns against any attempt to re-establish what it means to be free and rational.

Even though the postmodern position may seem neutral, viz. the contest between on the one hand Hobbes, Hume and Locke and on the other Hegel and Marx, its denial of the possibility of transcending the present (either temporally or theoretically) protects any current socio-economic status quo from a progressive discourse. Ryan (1982) writes: 'The turn against theory in the name of first-order narrative explanations is a part of the process by which such interests are shielded from rational critique.' To the extent that the status quo is defended by means of a narrative drawn from Hobbes, Hume and Locke, postmodernity is an unwilling ally of the latter.

However, there are elements of postmodernity which may enable a more sophisticated pursuit of the Marxist project hinted at in the previous section, if only by keeping Marxists on their toes. Too often in the past we tolerated blanket explanations of historical phenomena by those who hid their unsophistication behind certain linguistic forms. We were told that the reason why the working class in Britain failed to become a 'class for itself' was that the 'subjective' conditions were not there. Perhaps they were not, but this is not an explanatory theory. Imprisoned in the schematic metaphor of progress based on a unidimensional passage from changes in means of production to discrete changes in social relations, we tolerated the development of authoritarian regimes in Eastern Europe hoping that, eventually, the process of industrialisation would magically remove authoritarian political institutions and, thus, that socialist democracy would flourish.

Christopher Norris (1985) singles out the method of deconstruction of metaphors from the rest of the postmodern litany and argues that Marxists can find it useful. If what he means is that Marxists ought constantly to question the appropriateness of their metaphors, and to worry about the possibility that the metaphor may have been rendered inadmissible by some historical twist, he is correct. On the other hand, the work of Derrida has been done for him by Hegel and Marx. It was Hegel who focused on the contest between theory and narrative by criticising Kant for having elevated ideas to an a priori status when their true status is distinctly transient. And it was Marx who castigated the metanarrative dwelled upon by philosophers, arguing that there are no immobile absolutes, no spiritual beyonds and that every absolute represented a mask justifying exploitation of humans by humans. Although we do not need Derrida to tell us what Marx had

already elaborated (that is, that philosophical abstractions in themselves have no value or precise meaning), if Derrida incites us to return to the roots we neglected for so long, then all is well.

6.6 Praxis and the dialectic

The problems of truth-seeking and of defining liberty and rationality, have two things in common. First, they are seen as illegitimate by postmodernists, and second, they are solved simultaneously by Marx. Ironically, if postmodernity's contribution is simply to insist that transcendental solutions are illusory, it finds unexpected support from Marx who leaves little room for confusion when writing:

The question whether human thought can arrive at objective truth is not a theoretical but a practical question. It is in praxis that man must prove the truth, that is, the reality, the exactness, the power of his thinking. The dispute over the reality or non-reality of thinking isolated from praxis is a purely scholastic question.

[*Theses on Feurbach II*, in Marx (1964)]

The problem of knowledge in the abstract is a false problem. Abstract logical consistency, theory divorced from social activity and practical verification, have no value whatever. The essence of man is practical, and the essence of society is praxis – acts, courses of action, interaction. Separated from praxis, theory vainly comes to grips with falsely formulated or insoluble problems, bogs down in mysticism and mystification.

[*Theses on Feuerbach VIII*, in Marx (1964)]

Defining praxis as activity, viz. other humans, Marx seizes Hegel's argument, namely that Freedom and Reason are hollow in the absence of praxis, and refutes social theories which seek truth in a historical vacuum. When agents encounter social interactions (e.g.

Table 6.1

) which Aristotelian logic is unable to solve single-handed, they transcend their pre-political nature. The observation of historical change (and also of changes in cognition that occur in historical time) is central here. How does it come about if everything is self-referentially tied together without any external relation or foundation? There must be *something* which gives the process its dynamic. This something is the dialectic which we can simply think of as the foundational quality of society that renders it dynamic.

From this perspective, Reason becomes dialectical in that it is a concrete concept which does not live in separation from the history of human praxes. A simple way of conceptualising it is as a *process* of rationality-creation which codetermines the individual's motivation and the rules of rational choice. It is a process that is characterised as much by logic as it is by experiments with alternative rationalities.

16

And it is a process that neoclassical theory, loyal to its static spatial metaphors, cannot keep up with. Similarly for societies. As feudalism was about to vanish following the praxes of individual agents, suddenly it became intelligible as a socio-economic organisation and furnished us with the French revolution

concept of Liberty. At that point in history the meaning of freedom (and of Reason) was upgraded as a result of human social activity. Ever since, that meaning is constantly being threatened by ahistorical axiomatic definitions.

17

The above interpretation of dialectical Reason will not please many Marxists. For it entertains the possibility of an indeterminacy that only history can resolve. However, if the only way of undoing this indeterminacy *theoretically* (as opposed to *practically*) is to breach the canons of rational explanation (as game theorists do in the game of

Table 6.1

), then so be it. Note that this does not constitute a slide into postmodernity or pragmatism. The latter canvasses the non-availability of ultimate truths, the uselessness of large scale explanatory systems, and the view that ideas are interpretable only in terms of their past and present cultural relevance. At every point in time, the dialectical Reason proposed here is capable of grasping the logic implicit in those episodes of thought which make up its own pre-history. What it does not do is to specify *ex ante* the exact path on which it will tread. The dialectic's horizon is a state of perpetually

open possibility that human praxes determine in real time.

In this vein, the melange of praxes that constitute history becomes the determinant of meaning. On the level of the private, the dialectic describes the creation of perceptions and the actions that follow such perceptions. Those actions interact on the perceptions that caused them, give shape to the web of socially shared beliefs and are determined by it continually. Within this social framework, freedom is to be attained when social relations are in place which permit the Hegelian dialectic of recognition to function fully. But how can the self recognise the other when they meet at the capitalist market for labour in which one of the two loses her capacity to be a sovereign person? As long as social relations systematically restrict the options of certain groups or strata, the rest will be endowed with extractive power which ensures the impossibility of genuine freedom for the exploited *as well as for the exploiters*.

Of course we have all oppressed and we have all been oppressed at some stage of our lives. But the crucial point is the presence of systematic patterns of exploitation built into relatively primitive social relations. The structure of such social relations feeds into the constituents of unfreedom rendering constitutional, or axiomatic, liberty symbolic of what is unattainable under the existing socio-economic organisation. If Reason is the product of history, as Hegel would claim, then capitalism sets limits within which neither freedom nor rationality can breathe.

6.7 Epilogue: a functionalist explanation of the dance of the meta-axioms

This chapter examined the theoretical implications of the adoption by a certain group of Marxist scholars of a priori definitions of rationality and freedom which permeate neoclassical economics. I argued that this was an adoption that produced extremely thin analytical benefits at a gigantic cost to the adopters. Based on a Hegelian critique of ahistorical approaches to the meaning of Liberty and Reason, I have suggested that to preserve a connection with the reality of human rationality, freedom, as well as with the social world that we manufacture collectively on a daily basis, indeterminacy must be embraced as an irreducible aspect of both our freedom and our reasoning.

The problem with neoclassicism is that it tries to build a theory of capitalism first by postulating a determinate definition of free and rational agency and, then, once it hits the Wall of Indeterminacy that inevitably re-emerges (see

[Chapter 1](#)

), by recoiling behind logically illegitimate axioms. Marxists who were lured by neoclassicism to accept its definitions of free and rational agency were inevitably caught up in this vicious *dance of the meta-axioms*, ending up demoralised, lost and bewildered.

It is interesting to recall that this group of Marxists, RCMs as I call them, were motivated to seek 'enlightenment' in neoclassical game theory by the fatigue caused by the crudeness of functionalist explanations; of the type of 'lazy' explanation that many Marxists adopted as a matter of habit. While they accept that functionalist explanations can be 'good and proper', as in evolutionary biology for example, they concluded that social theory is best founded on intentional explanations. Let us, however, take a quick look at the prerequisites for functionalist explanation to make 'good and proper' sense in any scientific realm.

Functional explanations explain an organ, an institution, a phenomenon, by its beneficial effects. It is the effect of an action rather than an intention which lay behind the action which is used to explain why the action was taken. As Elster (1982, 1986b) himself argues, functionalist explanation is defensible, and useful, if it contains all five elements below (especially the fifth). The reason why RCMs rejected functionalism in the social sciences, and allowed themselves to be lured by game theory, was that they

judged that Marxist discourse usually relied on elements (1) to (4) but lacked element (5):

1. Y is an effect of X.
2. Y is beneficial for some agent Z.
3. Y is unintended.
4. The causal relation between X and Y is unrecognised.
5. Y maintains X by a causal feedback loop through Z.

Elster has argued that most functionalist arguments in social science (and particularly those in the Marxist tradition) fail to convince because they do not fill in how the unintended consequences of the action help promote the activity which is responsible for this set of unintended consequences. There has to be a *feedback mechanism*: that is, something akin to the principle of natural selection in biology which is capable of explaining behaviours by their 'success' and not by their 'intentions'. For the feedback mechanism to be present in any social theory, we need a demonstration of learning-through-adaptation. It is the assumption that people shift towards practices which secure better outcomes (without knowing quite why the practice works for the best) which is the feedback mechanism responsible for selection of practices. Thus in the debate over functional

explanation, 'learning' might supply the general feedback mechanism for the social sciences which will license functional explanations in exactly the same way as natural selection does in the biological sciences.

For example, suppose Y are the 'institutions of liberal democracy' (or 'sexism'), X is 'capital accumulation', and Z is 'capital'. For a functionalist explanation of liberal democratic institutions (or of sexism's staying power) to make sense, it is crucial not just to argue [element (2)] that it is beneficial to capital but also to demonstrate how [element (5)] liberal democratic processes (or sexism) helps maintain capital accumulation by means of a causal feedback process that operates through capital. If this fifth element is convincingly provided, then the functionalist social theory in hand has significant merits and cannot be dismissed as hocus-pocus.

It is clear that RCMs were somehow convinced that this fifth element would be provided by game theory. Alas, they were precisely wrong, for reasons that

Chapter 1

has explained in some detail. Interestingly, it is perfectly possible to take this argument further; to couch my *dance of the meta-axioms* narrative in terms of a proper functionalist explanation as follows: Let me cast the roles above in terms of

Z = neoclassical economists

X = the third meta-axiom

Y = neoclassicism's success as raising 'barriers to entry' that prevent non-neoclassicists from attaining discursive power in the public debate on how capitalism works.

In this vein, one might argue that neoclassical economists (Z) utilise assumptions such as that 'all agents' beliefs must always remain consistently aligned' (a form of the third meta-axiom) in order to close their models; so as to overcome indeterminacy and pinpoint the analytical solutions that will render their papers publishable in the renowned journals. However, unbeknownst to them, such an assumption comes wrapped up in a technical complexity that prevents non-neoclassicists (including well-meaning RCMs) from challenging the logical coherence of the resulting models. Thus, the act of recoiling behind the third meta-axiom (move X above) has the *unintended* consequence of erecting barriers to entry that leave neoclassicists unchallenged by non-neoclassicist critics, with greatly beneficial effects in terms of their discursive power – see Y above. Crucially, there is no conspiracy here. For this reinforcement of the neo-classicists' discursive power is wholly unintended because most neoclassical economists and game theorists (a) *believe* in the virtue of the equilibrium analysis founded on their consistently aligned beliefs assumption and (b) are not even aware of its function of

boosting their discursive power via keeping non-neoclassicists at bay! To conclude the argument, the unintended emergence of ‘barriers to entry’ (Y) proves beneficial for the ‘community’ of neoclassical game theorists (Z) in the competitive battle for resources within the academy and so maintains their position in the academy and reinforces their reliance on the third meta-axiom (X).

In other words, it is possible to explain the discursive success of neoclassical economics, even its capacity to magnetise a group of anti-functionalist Marxists, by means of a well founded... functionalist theory. In this, there is more than a hint of irony since RCMs, like Jon Elster, have championed game theory as an alternative to... functional arguments in social science. Well, if the recoiling behind neoclassicism’s third meta-axiom is essential in preserving game theory’s appeal, and if this is best explained functionally, then Marxists who turned to game theory in order to avoid functionalism ended up with a lot more (or is it a lot less?) than they bargained for!

Appendix 6.1: rational deliberation in the game of

Table 6.1

I have already presented a wordy version of how workers can rationalise a decision to rebel. The following table offers the formal proof that *all* strategies of each side can be supported by a rationalisable train of thought. W and C stand for the workers and the capitalists while b and denote the verbs ‘believes’ and ‘chooses’ respectively. Note that the equilibrium outcome occurs when workers and capitalists form thoughts A_2 and B_2 while the outcome (150, -50) follows thoughts A_3 and B_1 .

Workers’ choices and

their supporting thoughts

Strategy 1: $W \rightarrow 1$

because $WbC \rightarrow 3$

because $WbCbW \rightarrow 3$

because $WbCbWbC \rightarrow 1$

because $WbCbWbCbW \rightarrow 1$

A_1

Strategy 2: $W \rightarrow 2$

because $WbC \rightarrow 2$

because $WbCbW \rightarrow 2$

because $WbCbWbC \rightarrow 2$

because $WbCbWbCbW \rightarrow 2$

A_2

Strategy 3: $W \rightarrow 3$

because $WbC \rightarrow 1$

because $WbCbW \rightarrow 1$

because $WbCbWbC \rightarrow 3$

because $WbCbWbCbW \rightarrow 3$

A_3

Capitalists’ choices and

their supporting thoughts

Strategy 1: $C \rightarrow 1$

because $CbW \rightarrow 1$

because $CbWbC \rightarrow 3$

because $CbWbCbW \rightarrow 3$

because $CbWbCbWbC \rightarrow 1$

B_1

Strategy 2: $C \rightarrow 2$

because $CbW \rightarrow 2$

because $CbWbC \rightarrow 2$

because $CbWbCbW \rightarrow 2$

because $CbWbCbWbC \rightarrow 2$

B_2

Strategy 3: $C \rightarrow 3$

because $CbW \rightarrow 3$

because $CbWbC \rightarrow 1$

because $CbWbCbW \rightarrow 1$

because $CbWbCbWbC \rightarrow 3$

B_3

There is a view by some that there are two neoclassical theories on how to reach a solution. One is the equilibrium theory which I discuss here while the other is a Bayesian theory (for a relevant text see Skyrms 1990). The former seeks solutions that rational play *should* generate and then *assumes* that rational agents will assign a zero probability to any action by their opponents which does not comply with the equilibrium outcome. The excuse for this assumption is the assumption of rationality. The Bayesian approach is somewhat different. Initially agents are allowed to assign *any* subjective probability to the various actions of their opponents. Then, as they think about the game (or as the game progresses in repeated games), their prior beliefs are augmented through this process of rational deliberation until an equilibrium strategy is reached. Although these two theories do

not always yield the same result, there is an impressive degree of convergence to the same conclusion. This is so because of reliance on the same a priori definition of rationality. In the case of the equilibrium theory of games, the solution is arrived at because the theorist assumes that all *ex ante* expectations will be confirmed *ex post*, while in the Bayesian story there is in place a hidden assumption (often referred to as

the *Harsanyi doctrine* – see Aumann 1987) that agents are only allowed to do different things if they have different information or objectives. Thus both strands of neoclassicism banish the most important aspect of strategic uncertainty by imposing symmetry. By so doing, they smuggle in an implausible assumption to do the dirty work that their definition of rational deliberation is incapable of.

One wonders whether all neoclassical theorists are aware of this weakness of equilibrium analysis. I believe that here we have a brilliant example of motivated theoretical sloppiness. In a revealing passage by Robert Aumann (1987), a pillar of game theoretic orthodoxy, we find an acknowledgment that the symmetry axiom (i.e. the *Harsanyi doctrine*) is problematic. However, Aumann hastens to add that: ‘...economists feel that this kind of analysis [i.e. an analysis which recognises that rationality does not commend symmetry] is too inconclusive for practical use, and side-steps the major economic issues’. I can only infer from this honest statement that the theoretical problem is noted but ignored because it is too inconvenient!

Notes

- 1 Sugden (1990) shows how non-equilibrium choices can be construed as mutations that may prove to be evolutionary stable.
- 2 See for instance Martin Hollis’ (1987) description of the cunning with which Reason contrives to bind itself inextricably to social context once agents acquire a social dimension.
See his Introduction in (1986a) as well as chapters 1 and 8 of Elster (1986b).
- 3 I admit that this may be an unfair comment in that Wood (1990) has accused John Roemer for having his hand tied by ‘the narrowly formalistic requirements of the model and his subjection to the conceptual demands of neoclassical economics’. This sentence can be loosely interpreted to cover the criticism of rational choice theory advocated in this section. On the other hand, Wood (1990) identifies the problem with rational choice theory in its inability to specify ‘the social structures which set the terms of what is reasonable and preferable in any given context’. It is my view that rational choice theory, even if it has a perfect understanding of these structures, may be incapable of capturing the ‘reasonable and preferable’ choices. The purpose of my criticism is to augment, not to dispute, Wood’s criticism.
- 4 See, for example, Hargreaves-Heap (1989b) who tells a story of how agents involved in games with multiple equilibria generate historically contingent social conventions.
- 5 This is so in strategically interesting circumstance where no strategy of any player is dominated.
- 6 The usual rejoinder to any criticism of Rational Choice theory in general and game theory in particular is that there are refinements and extensions that the critic is unaware of and which accommodate the criticism. However, I believe that this constitutes an illusory escape route. In Varoufakis (1991) I show that every refinement or extension of equilibrium theory (e.g. the introduction of time and forward looking agents – in technical terms *subgame perfection*, asymmetric information and sequential or Bayesian equilibria etc.) is founded on the notion of equilibrium as outlined here. If the foundation is rotten, it takes more than ingenuity to build a robust explanatory structure.
- 7 Seyla Benhabib (1984) adds that for Hegel ‘from the standpoint of exchange no characteristic of individuals is relevant apart from the fact that each owns a certain property desired by the other’. There is an interesting parallel here with the so-called *Harsanyi doctrine* mentioned in [Appendix 6.1](#) in reference to the discussion in the previous section. In game theory, it is assumed that rational agents will behave in exactly the same way if they face the same payoffs and are fed the same information. This can be interpreted as an extension of the bourgeois/modernist assumption that it is only property endowments that make social actors different.
- 8 See Nozick (1974) for an example of the uses to which the Lockean proviso is put by modern libertarians.
- 9 That is, if market exchange is ‘...something embodying merely a common will and resulting from the arbitrariness of the parties united into a state’ [Hegel (1942)].
- 10 Hegel (1953) writes: ‘Rationality, taken generally and in the abstract, consists in the thorough-going unity of the universal and the single. Rationality, concrete in the state, consists (a) so far as its content is concerned, in the unity of the objective freedom ... and subjective freedom; and (b) so far as its form is concerned, in self-determining action on laws and principles which are thoughts so universal’. Hegel relates rationality and freedom as concrete mutually dependent notions that are intelligible only within an historical continuum.
- 11

That is, if the buyer purchases labour power and receives labour (whose value exceeds that of labour-power) because of the nature of exchange which is only partly market based.

12

There is a recent claim that history can be accommodated in the context of evolutionary game theory (EGT). EGT rejects the axiomatic approach to rational choice and substitutes it with quasi-Darwinian functionalism while retaining the assumption of frozen individual objectives. Its functionalism distances it from Rational Choice theory while its reliance on an unchanging human agency divorces it from history. See Chapter 8 of Varoufakis (1991) for more on the contest between the evolutionary and historical explanations.

13

In a famous passage Marx (1974) makes it clear that men and women produce their rationality as they create the rest of their lives.

14

In Varoufakis (1991b) I claim that it is possible to canvass a postmodern critique of game theory by showing that deconstructing the narrative of instrumental rationality may prove profitable for our players.

15

For instance, Kant assumes the ability to *know* things, Hegel our *consciousness* through the dialectic, Marx our *materially constituted* perception and Wittgenstein our ability to attach *meaning* to things.

16

Sugden (1990) shows how non-equilibrium choices can be construed as mutations that may prove to be evolutionary stable.

17

In a recent paper Van-Huyck *et al.* (1990) write: 'The power of the equilibrium method derives from its ability to abstract from the complicated dynamic process that induces equilibrium and to abstract from the historical accident that initiated the process.' Quite clearly, the ambition of game theorists is to rid themselves of history by distilling all the historically relevant information into their equilibria. My argument is that their equilibria are too analytically impoverished to bear the enormity of the task bestowed upon them by those who seek ahistorical explanations.

7 A theory of solidarity

Why indeterminacy is a prerequisite for genuine solidarity

7.1 Prologue

7.1.1 Background briefing

My modelling career (!) began in the early 1980s, at a time when Mrs Thatcher's programme for crashing organised labour was well under way. Involved, as I became, in a number of pickets in front of steel mills, demonstrations at Hyde Park, the Wapping 'experience' and, of course, the year-long miners' strike, I could not help but be utterly struck by the contrast between (a) the exceptional bonds of solidarity that I was witnessing on the streets, in working families' homes, in pubs, even along the high street, and (b) the pristine isolation typifying the 'life' of *homo economicus* (i.e. of the humanoid that lives in neoclassical economists' models).

When I would take a break from the tumult that was early 1980s Britain, for the purposes of creating my own economic models, the first task I set myself was to see if anything could be done to incorporate the notion of solidarity in economic theory. The models in

Chapter 2

were the result. In them I managed, to some extent, to infuse the missing 'ingredient' by introducing endogenous mobilisation variables that enabled the analyst to treat as important economic factors the social norms and conventions which allowed workers to overcome their prisoner dilemma tendency to defect from their common cause.

Alas, these models were a far cry from anything resembling real, authentic solidarity. Of that I remained painfully aware for years. When discussing the matter with philosopher Martin Hollis, a mentor as well as a University of East Anglia colleague, it became clear that defining solidarity analytically was a tall order. That unlike altruism, team-reasoning and such 'other'-regarding notions, solidarity was a slippery philosophical customer. Some time in the mid-90s I had the good fortune of spending some time at the Université Catholique de Louvain, at the invitation of Philippe van Parijs (who headed the Hoover Chair of Economic and Social Ethics). The purpose of my visit was to discuss and work on notions of solidarity. There I had the pleasure of meeting with Christian Arnsperger with whom we immediately struck up a friendship as well as a joint project to delineate solidarity analytically. The pages that complete this chapter is the result of that work.

¹ The reason why I decided to include this chapter here was none other than my conclusion on the matter, as expressed in this chapter's closing line: *Indeterminacy lies at the heart of authentic solidarity, just like it underpins good theatre, art and music.* Need I say more?

7.1.2 The rest of the chapter

While the consensus regarding the state's responsibility for sustaining the unfortunate and empowering the weak remained intact, the notion of *solidarity* invoked images of Polish dissidents and striking British miners. However, since the late 1970s the tide has been going out on many arguments in support of state-welfare systems. As it receded, the few weedy posts it left behind, especially on the run up to the crash of 2008, seem to have inspired a variety of European politicians and institutions

² to re-evoked solidarity, often as a means of counter-balancing the heightened emphasis on financialisation, entrepreneurship and self-reliance. However, it is not at all clear what calls for 'greater solidarity' could possibly mean. Is it a euphemism for organised philanthropy? For social constructivism funded by means other than taxation?

More likely than not, politicians and activists make use of the term because of its emotive value, with minimal clarity regarding what solidarity actually means. A good case in point is the appeal to 'solidarity' in order to justify the logically questionable 'bailouts' during the recent eurozone crisis. The analysis that follows in the chapter began with a query: Is solidarity a potentially useful analytical category? The result was an essay-in-retrieval on solidarity's *potential* meaning. It reflects the view that solidarity can be meaningfully distinguished from similar, and far better researched, other-regarding, dispositions; e.g. reciprocity, duty and altruism. Is solidarity just a sloppier term for what is already well-defined? Or does it open up a window to useful, fresh insights?

To begin with, let me postulate one basic prerequisite for solidarity; namely, a generous disposition; a propensity to sacrifice something one values (even if it only amounts to lost peace of mind) on behalf of some targeted group of people (e.g. refugees) whose welfare one deems important. Such generosity is defined formally in

[Section 7.2](#)

but nothing is said specifically on solidarity until

[Section 7.4](#)

[Section 7.3](#)

demonstrates that even *minimal* generosity, as long as it is commonly anticipated, can change the complexion of several classic social interactions (e.g. Rousseau's stag-hunt game). Six popular explanations of generosity are then discussed (ranging from natural sympathy and altruism to fairness equilibria) before solidarity is defined (see

[Section 7.4](#)

) as an analytically distinct other-regarding disposition.

[Section 7.5](#)

examines the special case of radical solidarity and links it to the evolution of arbitrary social power while

[Section 7.6](#)

offers the customary chapter epilogue.

[Section 7.4](#)

presents solidarity in juxtaposition to competing other-regarding notions. To give a flavour of the argument, I will be proposing that solidarity differs from altruism in that, whereas the latter is about treating the interests of other persons as one's own (or acting *as if* this were the case), solidarity is about identifying

a *condition* which makes those who 'suffer' it worthy of one's concern *independently* of (a) who those unfortunates are, (b) whether or not one cares for them *personally*. Put differently, altruism is a response to others' needs, interests and character. Solidarity, in contrast, is defined here as a reaction to a *condition* which afflicts certain 'others' independently of their personal or social character. And when this unfortunate *condition* is a product of social evolution, a social artefact in other words, then generosity turns radical and solidarity becomes subversive (see

[Section 7.5](#)

).

To the extent that my hypothesis is sensible and solidarity is, indeed, a form of targeted empathy toward strangers whose personal character is not the issue, it is considerably more puzzling than other forms of 'other-regarding' propensities. Unlike the conundrum of altruism, which has been addressed exhaustively in various ways,

³

solidarity-with-selected-strangers is almost as bewildering as Nietzsche's (1956) paradox of trust.

⁴

Of course, the analysis in this chapter does no more than to scratch the problem's

surface. At best, it opens up the debate and sets the scene for analytical treatments of a concept which is slowly re-gaining prominence in European political culture.

7.2 Generosity

All other-regarding deeds appear, at some level, as expressions of kindness or generosity. Thus, it seems natural to start our search for solidarity by postulating some minimal generosity that must characterise an act, or intention, before the latter is even considered as a possible expression of solidarity. Later I shall propose additional (sufficient) conditions which such acts must meet before specific cases of kindness can qualify as 'solidarity'; in juxtaposition to altruism, natural sympathy etc. So, let us begin with a simple definition of *perceived* generosity: We *believe* we are being generous to others if we act in a manner costly to ourselves but beneficial to them.

Suppose person i (who belongs to group M) is facing some choice problem and define S_i , $a_i \in S_i$, and $u_i(\cdot)$ as, respectively, i 's set of feasible actions or strategies, i 's chosen action, and i 's intertemporal utility function. Suppose further that, in i 's mind, there is a group of people, say N , who are affected by her choice. Then, the prerequisites of perceived generosity (see the previous paragraph) are in place if:

- (a) i 's choice $a_i \in S_i$ entails a sacrifice s_i for i , and
- (b) i thinks that group N members somehow benefited by her sacrifice s_i .

For action a_i to involve some sacrifice it cannot, by definition, be optimal from i 's own perspective. Thus, i 's optimal choice $a_i^* = \text{argmax}\{u_i(\cdot)\}$ must be different from her actual choice a_i and, thus, her sacrifice can be expressed in utility terms as $s_i = u_i(a_i^*) - u_i(a_i) > 0$. As for prerequisite (b) above, suppose that $W_N^i(a_i)$ is an index of group N 's welfare as perceived by i following i 's choice of $a_i \in S_i$. Then $w(a_i) = W_N^i(a_i) - W_N^i(a_i^*) \geq 0$ is an index of how much i thinks that

her sub-optimal choice a_i benefited group N .

Note that the usual aggregation problem does not apply here since the units of welfare utilised (w and $W_N^i(\cdot)$) represent no more than i 's *perceived* effect on the welfare of group N , as opposed to any real welfare effect.

In summary, the prerequisites for generosity, as stated above, take the form of simple inequalities: $s_i(a_i) > 0$ and $w(a_i) \geq 0$.

DEFINITION 7.1 Person i 's **λ -generosity** to members of target group N is given by

$$\lambda_i(a_i) = \begin{cases} S_i(a_i) \times w(a_i) & \text{if } S_i(a_i) > 0 \text{ and } w(a_i) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, person i (belonging to some group M ; $i = 1, \dots, M$) performs an act of λ -generosity toward members of group N ($j = 1, \dots, N$) if she acts in a manner which benefits them at her own expense. This definition is important for three reasons. First, it distinguishes sharply between generosity and reciprocity in the sense that, while acts of reciprocal kindness are underpinned by an expectation of something in return, the genuinely 'generous' are generous for nothing.

Secondly,

Definition 7.1

marks generosity out from its 'darker side'; namely, from spiteful acts intended at hurting others, at one's own cost (see note 6). Thirdly, because λ -generosity is the foundation upon which our solidarity concept is erected in Section 4. Before we can delve deeper into these issues, we need to say more about the beliefs in the background of λ -generous acts.

While generosity can be random and lack reasons, to qualify as something more 'substantive' (e.g. as justice in action, or solidarity, or team-reasoning) actions must be

grounded on specific reasons.

8

To begin with, for $i \in M$ and $j \in N$ we let $\lambda'_{ij} = E^j[\lambda_i]$ denote the expectation of j regarding i 's λ -generosity to her group and $\lambda''_{ij} = E^i[\lambda'_{ji}]$ i 's estimate of j as i 's expectation of λ_i . For example, when $\lambda_i > 0$ and $\lambda''_{ij} > 0$, i intends to be generous to N -members and thinks that this is precisely what they expect her to do. The rationale here is that other-regarding acts are often driven by the power of others' expectations.

9

And when one's 'sacrifice' is directed at a whole group (N), then average expectations among that group, as well as one's assessment of what people in a similar position might do (fellow M -members), play a crucial role in capturing our agent's situation. So, to complete a profile of other-regarding actions and beliefs, we define $\Lambda_{M \sim i}$ as i 's expectation of the average λ_i that others like her (i.e. also belonging to group M) will choose (or would have chosen under similar circumstances) and Λ_{Nj} as the average λ_i i anticipates members of the target group N expect (on average) of M -members like her.

$$\Lambda_{M \sim i} = E^i \left[\frac{1}{M-1} \sum_{i \neq k=1}^{M-1} \lambda_k \right]; \Lambda_{Nj} = \frac{1}{N} \sum_{j=1}^N \lambda''_{ij}$$

So far we have looked at i 's calculative second-order predictions,

10

viz. members of both her group and of those she wishes to benefit. Typically though there is another type of belief that plays a significant role in motivating agents: normative beliefs. To introduce such beliefs in i 's deliberations, we define ξ_i as i 's belief about the value of λ_i that she *ought* to choose; ξ''_{ij} as i 's (predictive) belief about what j ($\in N$) believes λ_i *ought* to be,

11

and $\Xi_{M \sim i}$ and Ξ_{Nj} as i 's expectation of average opinion regarding the value of λ_i that she *ought* to choose amongst her fellow M -members as well as N -members respectively. That is,

$$\Xi_{M \sim i} = E^i \left[\frac{1}{M-1} \sum_{i \neq k=1}^{M-1} \xi''_{ik} \right]; \Xi_{Nj} = E^i \left[\frac{1}{N} \sum_{j=1}^N \xi''_{ij} \right]$$

DEFINITION 7.2 Agent i 's λ -profile is given by $\langle \lambda_i(a_i) | \{\Lambda_{M \sim i}, \Lambda_{Nj}\}, \{\xi_i, \Xi_{M \sim i}, \Xi_{Nj}\} \rangle$

In brief, a person's λ -profile is defined by (a) her λ -generosity toward members of group N , as conditioned on, (b) her calculative (the Λ s), and (c) her normative beliefs (the Ξ s). To illustrate, suppose $\lambda_i = \xi_i = \Xi_{Nj} > 0$ while $\Lambda_{Ni} = \Lambda_{M \sim i} = \Xi_{M \sim i} = 0$. In this case, i makes a sacrifice which she expects to have positive effects on the welfare of some target group N ; she believes that she *ought* to be making such a sacrifice (and that N -members think so too); she also thinks that no one predicts that she would in fact prove so generous. Indeed, she is of the view that fellow M -members dismiss any notion that she is morally obliged to make sacrifices on the behalf of group N .

7.3 The impact and sources of generosity

So far, nothing has been said pertaining exclusively to solidarity. Indeed, the definition of an agent's λ -profile above may be helpful in depicting, and dissecting, all sorts of other-regarding behaviour, including altruism, or even love. While our particular hypothesis on what distinguishes solidarity from related concepts will have to wait until the

Section 7.4

, it might be useful to emphasise one impression that the word 'solidarity' conjures up: solidarity, by nature, involves large numbers of people. In contrast, love and altruism

seem to be better suited (though not exclusively so) to small groups.

The object of one's romantic love is, usually, a sole person. Altruism *may* be confined to a mother's feelings towards her offspring. Collusion usually involves no more than a handful of agents. By comparison, it seems harder to envision solidarity in a similar context; for it usually entails a generosity of spirit that extends to larger numbers, in which love and altruism have a tendency to dissolve. Coal miners caught up in some underground emergency are more likely to expect of their colleagues a degree of solidarity, or team-reasoning, rather than love, pure altruism or the type of reciprocal logic that motivates collective action against the employer.

Of course these are just preliminary thoughts which we shall return to after our attempt at a definition of solidarity in

Section 7.4

. Meanwhile, it is interesting to examine one common thread running through different types of other-regarding motivations which, like solidarity, are more relevant when more than two people are involved (e.g. reciprocity, norm-driven behaviour, team-reasoning etc.): The common thread in question is the thought that such other-regarding behaviour toward, as well as within, some target group N is inextricably linked to the group's shared identity. Moreover, a shared identity allows agents to coalesce to the *common* expectations of group N , or to *knowledge* regarding their generous disposition toward them. And when this happens, as we shall see below, some interesting results follow.

DEFINITION 7.3 Commonly known λ -generosity (CKG_{λ}^*) toward group N requires that, (a) each agent $i \in M$ chooses a sacrifice level s_i at least equal to $s^* > 0$ for the purpose of boosting the welfare of group N by $w^* > 0$; (b) knows that all other agents $j (\neq i) \in N$ know (a); (c) all agents $j (\neq i) \in N$ know (b); *ad infinitum*. By definition, $\Delta_{Ni} = \Delta_{M \sim i} = \lambda^* = s^* \times w^*$, while the normative expectations $(\xi_i, \Xi_{Nj}, \Xi_{M \sim i})$ could differ from λ^* . When, however, $\xi_i = \Xi_{Nj} = \Xi_{M \sim i} = \lambda^*$, we have a stronger case of CKG_{λ}^* in the sense that the agents' calculative beliefs are reinforced by (identical) normative ones.

A vivid illustration of the analytical value of CKG_{λ}^* can be given in the context of a simple interaction in which an infinitesimal λ can solve a perennial problem in game theory, as long as it is commonly known. By *minimal generosity* we shall henceforth refer to a case of CKG_{λ}^* with $s^* = \varepsilon$, where ε is vanishingly small but never zero. Consider the following one-shot game in which, for simplicity, sets N and M coincide: Suppose each person $i \in N (= M)$ must choose a real number a_i from the interval $[1, 10]$. The payoff function for each player is: $u_i(a_i) = A \times \min(a_i, a_j) - a_i \quad \forall i, j (i \neq j) \in \{1, \dots, N\}$ where $A \neq 1$.

Clearly this game is of the N -person coordination-problem type (also known as the stag-hunt game, see note 11) featuring an infinity of Pareto-ranked Nash equilibria within the continuum $[1, 10]$. Ultimately everyone is best off when each chooses $a_i = 10$ [in which case $u_i = 10(A - 1) \geq 0$] and no player has an incentive to select a number below that chosen by others. Nevertheless the Nash best reply strategy is to choose the smallest number in $[1, 10]$ that one predicts will be selected by anyone within the group [i.e. set $a_i = m$ where $m = E^i\{\min(a_j)\}, \forall i, j \in N$]. Thus, even the slightest degree of pessimism (i.e. $m < 10$) suffices to lead players to an inefficient outcome. Indeed experimental work has shown that, often, the greater the experience of subjects with this game the lower their payoffs.

¹²

Instrumental rationality, even when commonly known, cannot guarantee successful coordination in this game despite the absence of inbuilt incentives to 'cheat' or 'defect'.

¹³

However if, additionally, players act under common knowledge of *minimal*

generosity, successful coordination on the Pareto-dominant Nash equilibrium is guaranteed. To see why, suppose that, in equilibrium, each player

expects a Pareto-dominated Nash equilibrium with everyone in the group choosing $a_i = \alpha (< 10) \forall i \in N$. *Minimal generosity* means that each player will be prepared to make a tiny sacrifice $s_i = \epsilon$, an act of λ -generosity, in order to benefit the rest and will thus choose $a_i = \alpha + \epsilon$ (Nb. It is easy to show that if $s_i = s^* = \epsilon$ then $a_i = m + \epsilon$). At this stage, we have $\langle \lambda_i(a_i) = \lambda' > 0 | \{\Lambda_{M \sim i} = \Lambda_{Nj} = 0\} \rangle$ where $\lambda' = \epsilon(N - 1)(A - 1)$.

14

But then, courtesy of *minimal generosity*, everyone will anticipate i 's new profile $\langle \lambda_i(a_i) = \lambda' > 0 | \{\Lambda_{M \sim i} = \Lambda_{Nj} = \lambda'\} \rangle$ and thus their estimates of α will be revised upwards. All of a sudden the ϵ -increase in a_i is no longer an act of sacrifice, or λ -generosity, since choice $\alpha + \epsilon$ is a Nash best reply strategy to the new expectations. In other words, the agent's profile is transformed again to $\langle \lambda_i(a_i + \epsilon) = 0 | \{\Lambda_{M \sim i} = \Lambda_{Nj} = 0\} \rangle$. Since by this stage of their deliberation no generosity is required, to be minimally generous is to choose $\alpha + 2\epsilon$; that is, each player's λ -profile is revised upwards to $\langle \lambda_i(a_i = \alpha + 2\epsilon) > 0 | \{\Lambda_{M \sim i} = \Lambda_{Nj} = 0\} \rangle$. And so on, until each player's λ -profile becomes $\langle \lambda_i(a_i = 10) = 0 | \{\Lambda_{M \sim i} = \Lambda_{Nj} = 0\} \rangle$. At that point all choose $a_i = 10$, the Pareto-dominant Nash equilibrium is achieved, and no *actual* generosity is necessary.

15

To sum up, once minimal generosity is taken for granted by all, coordination is achieved without any need for mutual sacrifices. This interesting result can in fact be generalised for a class of continuous, finite N -player coordination games.

PROPOSITION 7.1 *In N -person interactions with continuous strategy/payoff spaces, multiple Pareto-ranked Nash equilibria, and risk-dominance of the Pareto inferior equilibria (over the Pareto-superior ones), the Pareto-dominant mutual-maximum (Nash) outcome will occur if players act under common knowledge of (a) minimal generosity (**minimalCKG***) and (b) instrumental rationality (**CKIR**). Moreover, no generosity will be shown in equilibrium.*

Proof: I consider games in which each player's strategy a_i is chosen from a continuous, closed and bounded set $S_i \subseteq \mathbb{R}$ with a common upper bound (\bar{a}). Further, the players' payoff functions u_i are also continuous mappings such that: (i) the game features multiple Pareto-ranked Nash equilibria; i.e. $u_i(a_i = a) > u_i(a_i = a - \epsilon) \forall i, j \in N$ and $\epsilon > 0$; and (ii) no player has a capacity to increase her payoffs by choosing an a_i below the smallest choice in the group; that is, i 's best reply to the expectation that the smallest choice will equal m [i.e. $m = E\{\inf(a_j)\}, \forall i, j \in N$] is to select strategy $a_i = m$. By definition, when everyone selects \bar{a} as their strategy, each collects the highest available payoff; a mutual-maximum equilibrium: $u_i(a_j = a) \geq u_i(a_j \leq a) \forall i, j \in N$. In this equilibrium, private and social optimisation is achieved and, by our earlier definition of λ -generosity, no agent gets a chance to put their generosity on display since each chooses the behaviour that serves their narrow self-interest. Under the assumption of a continuous strategy/utility space, it is evident that no other symmetrical outcome (i.e. a strategy choice of $a_j < \bar{a}, \forall j \in N$) is consistent with both Nash equilibrium and *minimal generosity*. To see this, suppose that each player is contemplating strategy $a_j < \bar{a}$ and everyone knows this. Under *minimal generosity*, each expects everyone else

to be ready to make a slight sacrifice on behalf of the rest; that is choose $a_j + \epsilon$ instead of a_j . Due to the continuity assumption, a new Nash equilibrium exists in pure strategies: players choose $a_j + \epsilon \forall j \in N$ and thus anticipate a uniform rise in their payoffs. Once this stage in the iterative process is reached, agents again optimise (and their sacrifice level returns to zero). A new iteration therefore starts as *minimal*

generosity, once more, motivates players to revise their strategy upwards to $a_i + 2\varepsilon$. And so on until the iterative process reaches its upper barrier at the mutual-maximum equilibrium at which, as shown above, actual generosity is neither necessary nor possible. \square

The interesting feature of the above result is that *generosity*, even in tiny doses, succeeds (as long as it is commonly known) where hyper-rationality has hitherto failed: in procuring an all-round beneficial (that is, Pareto superior) equilibrium. As long as the players' payoffs are continuous functions defined in a continuous strategy space, even infinitesimal values of ε will gradually dispel pessimistic expectations and push players' strategies in the direction of the mutually most beneficial equilibrium. The above proposition is of course relevant for a fairly narrow class of social interactions: Continuous coordination games in which i 's a higher than average contribution (or sacrifice s_i) benefits the other player(s) (however infinitesimally). *Minimal generosity* suffices in such games because Jill has an opportunity to be minimally λ -generous to Jack in every Nash equilibrium (and vice versa). This is the hook that the algorithm requires to generate full coordination out of minimal solidarity.

By contrast, no such hook is available either in pure coordination problems or in antagonistic games (e.g. such as hawk–dove, prisoner's dilemma). In the former case (i.e. pure coordination), once they have homed in on some equilibrium (however Pareto inferior it might be), agents have no way of making the requisite minuscule sacrifice on behalf of fellow players. Similarly, in the case of antagonistic games; once a conflict of interest emerges (e.g. when different equilibria are favoured by different people or players have clear incentives to 'defect', as in the prisoner's dilemma), *minimal generosity* fails to make a difference. In those richer contexts we shall need to examine the connection between a person's degree of λ -generosity and the underlying beliefs within her λ -profile. Nevertheless, it was still rather important to have shown (see above) that there *does* exist a class of social interactions in which even minimal, commonly known, generosity can forge hearty bonds between atomistic individuals. The question now is: What motives underpin commonly anticipated generosity? In the remainder of this section I review a number of well-researched sources of such motives. In the next I argue that solidarity is quite distinct from these and deserves to be treated as a separate notion.

(a) *Team-reasoning*. According to Sugden (1993) and Bacharach (1999) individually rational persons sometimes manage to see themselves as members of a team whose common purpose bears significantly upon their private passions. When this happens, a general commitment to the team's objectives is taken

for granted and various coordination difficulties disappear. Precisely the same point was made in the previous section; namely that several coordination failures are avoided once agents are embroiled in *minimal generosity* (or **minimal CKG $^*_\lambda$**). Thus, *team-reasoning* and *minimal generosity* are analytically equivalent. They help resolve the same class of coordination problems while, at the same time, they fail in equal measure to foster cooperation at the slightest hint of conflicting interests between agents. For example, in interactions of the prisoner's dilemma type, *team-reasoning* dissolves in the wake of the centrifugal forces created by private agendas and *minimal generosity* is too brittle to overcome the destructive logic of free-riding. Something more is needed. Indeed in the context of a free-rider problem (or N -person prisoner's dilemma) that 'something' is *maximal generosity* (or **maximal CKG $^*_\lambda$**).

DEFINITION 7.4 Maximal generosity toward group N requires commonly known λ -generosity (**CKG $^*_\lambda$**) among members of group M with $\lambda^* = \lambda_i(a_m)$ for each $i \in M$ and $a_m = \arg\max_{a_i}(\lambda_i(a_i))$.

Example: Consider a free-rider variant of the earlier N -person interaction. Each player

selects a real number in the $[1, 10]$ interval and receives payoffs:

$$u_i(a_i) = A \times \frac{1}{N} \sum_{j=1}^{N=M} a_j - a_i \quad \forall i, j \in N(=M), \text{ with } N > 1 \text{ and } A \neq 0$$

16

In the previous game *minimal generosity* guided instrumentally rational agents safely to the mutually maximum outcome. In this free-rider version, however, the dominant strategy is to choose 1 regardless of what the others will do and, therefore, nothing less than a (commonly known) readiness to be maximally λ -generous (i.e. choose 10 rather than 1) will do the trick.

17

PROPOSITION 7.2 *In free-rider/prisoner dilemma games, the mutual-maximum (non-Nash) outcome will be selected if players act under common knowledge of maximal generosity (**maximal**^{CKG*} _{λ}) given their beliefs regarding their opponents' choices.*

Proof. Consider the simple two-person prisoner's dilemma in which each player chooses between strategies 'defect' (d) and 'cooperate' (c) and faces the following utility preference ordering $u_i(d, c) > u_i(c, c) > u_i(d, d) > u_i(c, d)$ for $i = 1, 2$; where $u_i(a, b)$ is i 's utility from playing strategy a while the other player chooses b . By virtue of strict dominance, their optimal action a_i^* is to select strategy d independently of their expectations. To do otherwise (i.e. to select $a_i \neq a_i^*$) requires *maximal λ -generosity*: If 1 expects 2 to choose her dominant strategy d , in choosing c agent 1 is selecting the maximum sacrifice s_1 possible $\{s_1 = u_1(c, d) - u_1(d, d)\}$ and the largest welfare benefit to her opponent $\{w = u_2(d, c) - u_2(d, d)\}$. If, on the other hand, 1 expects 2 also to be λ -generous, that is to play strategy d , in choosing c player 1 is selecting the sacrifice level $s_1 = u_1(c, c) - u_1(d, c)$ and estimates the welfare benefit to her opponent as $w = u_2(c, c) - u_2(c, d)$. In the special case where

$u_i(d, c) - u_i(c, c) = u_i(d, d) - u_i(c, d)$, the degree of λ -generosity necessary to bring about a cooperative action is maximal and independent of the actor's beliefs regarding her opponent's intentions. When this equality does not hold, then a necessary and sufficient condition for cooperative moves is that players adopt maximal λ -generosity given their beliefs about the opponent's move. A similar result holds in N -player versions of the game. For instance, in the free-rider game above, any strategy

$$s_i(a) = \left(1 - \frac{A}{N}\right)(a - 1)$$

choice $a_i = a > 1$ corresponds to a sacrifice level equal to while the welfare impact of such λ -generosity to the remaining $(N - 1)$ players equals

$$w = \left(\frac{A}{N}\right)(a - 1)(N - 1)$$

. Thus, the precise level of λ -generosity by player i (whenever she strays from her dominant strategy $a_i = 1$) is given as

$$a_i = 1 \text{ is given as } \lambda_i = s_i \times w = A(N - 1)(N - A) \left(\frac{a - 1}{N}\right)^2$$

. In this case, due to the linearity of the payoffs, it is clear that a player's λ -generosity is independent of her beliefs regarding how others will behave. Moreover, to reach the decision to play in a fully cooperative manner (that is, set $a_i = 10$), we require maximal λ -generosity.

18

When payoff functions are non-linear, again we require maximal λ -generosity, only this time the latter will vary with the players' beliefs about their opponents' behaviour. \square

Some authors have argued, controversially (see, for example, Gauthier, 1985) that, in the context of free-rider games, any level of λ -generosity below its maximal value is an instrumentally irrational choice. Their point is that it would be profitable to develop a *disposition* toward *conditional cooperation*, which in our terms translates into arguing that there are good instrumental reasons for cultivating in our hearts and souls an

λ -profile which comprises maximal λ values as long as Λ_{M^i} and Λ_{N_j} exceed some threshold. However this is an unconvincing argument because, at least in one-shot free-rider interactions, values of λ_i significantly greater than zero cannot be explained unless agents are motivated by something beyond an urge to increase their direct utility.

19

Below we examine well-known suggestions as to what that ‘something’ might be.

(b) *Hume’s natural sympathy, Smith’s moral sentiments and utilitarian altruism*: Moved by sympathy, the ‘chief cause’ of moral practice according to David Hume, the agent may think of others’ interests as her own (though in inverse proportion to the psychological distance between her and ‘them’).

20

Similarly with generosity occasioned by Adam Smith’s moral sentiments.

21

Given sufficient sympathy or sentiments for members of group N , the value of λ chosen by a Humean/Smithian can be quite substantial. On the other hand, the fact that neither sympathy nor sentiments extend to all people and all groups is what creates the need for, and the possibility of, justice. To be just is to be generous to those for whom one harbours no ‘natural sympathy’ or ‘moral sentiments.’ Though not necessarily an end in itself, pleasure derives from acting justly

toward others; something that can only imply that a sacrifice was made on their behalf at odds with one’s narrow self (or family, or class) interest. ‘With regard to all... benevolent and social affections’, wrote Smith (1976), ‘it is agreeable to see the sense of duty employed rather to restrain than to enliven them, rather to hinder us from doing too much, than to prompt us to do what we ought. It gives us pleasure to see a father obliged to check his own fondness, a friend obliged to set bounds to his natural generosity, a person who has received a benefit, obliged to restrain the too sanguine gratitude of his own temper.’

Utilitarians have a simple explanation of positive sacrifices $s_i > 0$ on the behalf of target groups. Having reduced all of the agent’s passions (including her natural sympathy to others) to a single one (i.e. the maximisation of utility function u_i),

22

positive s_i values stem from an inner cost-benefit analysis. To be precise, an altruistic act a_i^o , involving sacrifice level $s_i > 0$, is performed when $a_i^o = \operatorname{argmax}_i \{u[s(a_i), w(a_i)]\}$; i.e. because this sacrifice leaves the agent at a higher point of her scale of ordinal preference. In this case, both the coordination and the free rider problems (examined above) recede in proportion to the valuation of others’ welfare (i.e. to $\partial u/\partial w$).

23

However, we note that such sacrifices do not qualify automatically as cases of λ -generosity – recall *Definition 1* and its insistence that generosity must involve a loss of net utility. However, utilitarians may get around this requirement by distinguishing between direct and indirect utility; namely, between utility that does not take into account the psychological benefits from having acted selflessly and utility that does.

(c) *Kantian, rule-utilitarian and Rawlsian generosity*: A Kantian propensity to be generous is independent of any pleasure that might be derived from it. Generosity, of this ilk, is a matter of doing one’s ‘duty’; and, in Kant’s (1949) infamous words, ‘the majesty of duty has nothing to do with the enjoyment of life.’ In the same way that the Kantian is duty-bound not to break a promise (since she cannot will that everyone should break theirs), our Kantian refuses to set her λ s equal to zero (even when her net utility suffers as a result). Thus, Kantians are, by construction, maximally solidaristic. In both games thus far examined, Kantians set $a_i = 10$ even though they are fully aware that smaller choices (i.e. contributions to the social group) are individually more lucrative. For Kant has defined rationality as a capacity to overcome the temptations of hypothetical reasoning and to stick to its categorical variant which enables, indeed forces, the rational person to recognise her duty to do what is right as opposed to what

is expedient.

24

Rule-utilitarians follow a similar, but quite distinct, logic. They ask: 'What degree of generosity would maximise my utility were it to be chosen by all, including myself?' Again the unique answer in both relevant games is to select, as part

of a rule or a disposition, the maximal sacrifice. Interestingly, both Kantians and rule-utilitarians end up with higher payoffs (e.g. as a result of successful coordination and/or cooperation). But rather than being the *reason* for their generosity, this welfare improvement is merely a satisfying by-product.

To recap, a Kantian's λ -generosity makes itself felt in the form of sacrifices performed in the *line of duty*; that is, independently of any cost-benefit calculation and unmoved by the expectations of others. It is in this sense that a Kantian's minimum

25

level of λ_i is always independent of the other arguments $(\xi_i, \Xi_{Nj}, \Xi_{M\sim i}, \Lambda_{Ni}, \Lambda_{M\sim i})$ in her λ -profile. Rule-utilitarians are less high-minded than Kantians (as utility is their ultimate guiding force) and more generous than straightforward utilitarians (since, unlike the latter, they are capable of generosity *as a rule*).

An analytically equivalent interpretation of *maximal generosity* can be attained by invoking Rawls' (1971) veil of ignorance. It is akin to a willingness, by an agent belonging to group M , to select an action after imagining that, *ex post*, one will end up either as still a member of group M or of another, less fortunate, group N (without knowing *ex ante* which of those M or N people one will turn into). If that 'blind' choice were to be made under the influence of infinite risk aversion, the resulting λ -generosity would equal $\lambda_i \equiv \xi_i = \max_{\lambda_i \equiv \xi_i} [\min_{k \in M \cup N} u_k]$, irrespective of i 's expectations.

26

(d) *Conformity with others' predictive beliefs*: Olson (1965) makes the obvious point that persons are motivated by an urge to 'win prestige' amongst their peers. Becker (1974) adds the fear of being scorned. Such motivation would lead an agent to select λ_i in proportion to Λ_{Nj} and/or Λ_{Mi} because when, say, Λ_{Mi} is high she loses utility if *seen* to act selfishly (i.e. if seen to choose $\lambda = 0$). Akerlof (1980) produced a dynamic version of this story by modelling the relative weight of Λ_{Nj} in one's utility as an increasing function of Λ_{Mi} . In other words, as long as a minimum level of sacrifice (or λ -generosity) is anticipated, then a bandwagon effect begins to unfold and 'selfless' acts spread inexorably.

27

More recently, Brennan and Pettit (2000) extend these ideas in their study of the urge to cultivate esteem.

Geanakoplos, Stacchetti and Pearce (1989) and Sugden (2000) delve deeper in suggesting a direct link between beliefs and preferences. They model an agent's preferences as a direct function of her second-order beliefs; that is, an agent might prefer to act in solidarity with group N , *even if no one is to know*, as long as she thinks that this is what is expected of her. To see how this idea differs fundamentally from Olson (1965) and Becker (1974), consider two examples. First, in the models by Olson and Becker, if my actions are unobservable by others then there is nothing that would motivate me to be generous. Invisibility would remove the lure of prestige acquisition or the threat of losing face. However, in Geanakoplos *et al.* (1989) and Sugden (2000) the mere fact that some people *expect* me to make a sacrifice makes me *want* to make that sacrifice (irrespectively

of whether I am being monitored or not). Secondly, Geanakoplos *et al.* (1989) allow for the possibility that agents who act on these reasons might, nonetheless, regret the fact that others entertain 'great' expectations of them; a case of what we might term

reluctant generosity.

28

(e) *Conformity with others' normative beliefs*: This is a variant of (d) above with others' normative beliefs replacing their calculative ones in *i*'s λ -*profile*. Once more, others gain a hold on one's utility, either directly or indirectly, as their moral beliefs influence the agent's preferences. Of course, the moment predictive beliefs are 'allowed' to contaminate preferences (e.g. Geanakoplos *et al.*, 1989, Sugden, 2000 and

Chapter 8

below), the distinction between positive and normative beliefs becomes really fine. If one's behaviour is influenced by an urge not to frustrate others' beliefs, and this is common knowledge, beliefs appear simultaneously as predictive and normative. Nevertheless, we think that the appearance of a fully collapsed distinction is deceptive. It is one thing to help a needy person because others *predict* you will do so (and know that their predictions matter to you), and it is quite another to help because, otherwise, that they would think of you as morally defective.

(f) *'Biblical' generosity*: Imagine that person *i* plans to make sacrifices for group *N* because she thinks that, had *they* been in *her* shoes, they would be prepared to make similar sacrifices. Note a crucial difference between this and straightforward utilitarian reciprocity (which we have referred to previously as *enlightened selfishness*). In the latter case you help others because the expected benefits are significant (e.g. tit-for-tat cooperation in a repeated free-rider game). The same applies, though at the level of the unconscious, to or socio-biological reciprocity. Here, however, we are referring to a different motivation altogether: An agent *i* is prepared to act selflessly, and at a cost, *independently* of any *actual* benefits to be had from such action. The mere thought that group *N* members are well-disposed to her, that they would have helped her if they had swapped places, is sufficient reason to want to help them even if she thinks it impossible that such a reversal of fortune will occur. In this sense, *i*'s λ -*generosity* will be positive regardless of whether she expects to benefit materially from it. It is intentions that count alone and, therefore, such beliefs can potentially lead to positive λ -*generosity* even in one-shot free-rider interactions.

However, this type of generosity has a nasty underbelly and it is for this reason that we use the term *biblical* to describe it. The ugly flipside transpires when we consider the possibility that *M*-group members fear that their *N*-group counterparts would be willing, if they could, to make positive sacrifices (s_i) in order to harm them. As a result, they are motivated also to make positive sacrifices to hurt them back. Indeed when both groups feel the same way about one another, we may end up in equilibrium with positive s_i values, negative welfare effects w_i , and no λ -*generosity* (since the latter is zero under these circumstances even if product $s_i w_i$ is non-zero).

29

A generalisation of this idea

allows for the possibility that cohesion and mutual generosity *within* one group (*M*) might well be dependent either on the mutual generosity or hostility with another (*N*).

30

7.4 Solidarity

The last section examined six other-regarding categories of generosity. In this section I propose that *solidarity* should be added to these as a distinct analytical category of other-regarding motives and acts. To demonstrate why I think this, let us re-visit Sugden's (1993) example of the *British Lifeboat Service*; an institution financed entirely through public donations. 'Why do people contribute money to it?' asks Sugden. He points out that the answer cannot lie in utilitarian altruism. For if donors are motivated by an interest in ensuring that the Service has sufficient funds to perform its lifesaving duties, they ought to think of each contributed pound as a perfect substitute for each pound contributed by someone else. Yet the econometric evidence contradicts this

hypothesis.

31

Selten and Ockenfels (1998) make a similar point. They report that, in an experimental setting, winners of a simple lottery proved quite willing to donate a portion of their winnings to the losers but, surprisingly, their donations turned out to be largely *independent* of how much the latter collected from other donors, or even of how the donations were to be divided amongst a number of recipients.

32

This result, just like the econometric evidence reported in Sugden (1993), amounts to a violation of utilitarian altruism's requirement that donors' valuations of recipients' utility from contributions be symmetrical vis-à-vis the contributors.

33

In both examples, donors are channelling their empathy to a particular target group (e.g. the 'shipwrecked', the 'lottery losers'). The question is: On what basis is this group selected? The usual explanations turn on (a) personal characteristics and (b) universalisable principles. We are generous to persons from whom we expect something back (even if it is only their gratitude); who belong to the same team/group as we; for whom we care individually; or toward whom we have a sense of universalisable duty. But this chapter attempts to highlight a different motivation: We may be generous to a class of persons (even when none of the above apply) simply *because we identify with their condition*. The resulting definition of *solidarity*, below, draws on this capacity.

Before proceeding further with the definition of solidarity, it is important to note that solidarity may, of course, coexist with reciprocity,

34

person-specific sympathy, team-reasoning and Kantian duty. The point, however, is that solidarity motivates generosity *independently* (that is, even in the absence) of these other-regarding motivations. The source of its power comes from nothing more than the fact that these are people unwittingly connected by some shared condition (e.g. ship-wrecked, HIV-infected) which fuels our solidarity toward *whoever* might be afflicted by it. Therefore, we envisage solidarity as a *condition-specific* disposition.

Given that solidarity (as defined here) does not rely on the expectation of reciprocal generosity, and in view of its impersonal (and condition-specific) nature,

it is obvious that solidarity cannot be a species of 'enlightened selfishness' or utilitarian altruism.

35

The same applies to team-reasoning and Kantian duty, neither of which explain this aspect of human motivation. The reasons for thinking this follow:

Team reasoning requires team spirit and, by definition, excludes all acts of solidarity by non-members. Though there is no hard evidence on this, it seems likely that a large part of the funds received by the Lifeboat Service come from non-sailors. Why would, for instance, a poor land-bound single mother give money to support a sea-rescue service? It seems far-fetched to suggest that her motivation is tantamount to natural sympathy or altruism toward rich round-the-world yachtsmen with more money than sense. Nor is it plausible that she fancies herself as part of their jet-setting 'team'. However, she may well contribute if she feels that the shipwrecked are *entitled* to *her* help in virtue of being shipwrecked and independently of who they are or how much others help them. Similarly, with the winners in Selten and Ockenfels (1998). Given the experimental design, it is hard to imagine that subjects managed to develop in the laboratory the bonds which occasion team-reasoning. It is more credible to suggest that the winners donated money to losers, not because of some concern about how much money fellow players leave the laboratory with, nor because winners feel they belong to the same group as losers, but due to a feeling of solidarity with the losers *as losers*; a feeling which breeds an obligation to share with them part of one's winnings.

Why is this obligation not some form of Kantianism? Kantians are λ -generous because they *ought to*, even if they feel no empathy with the person afflicted by the condition that gives rise to their duty. They are capable of donating to the Lifeboat Service (independently of their feelings toward sailors) because of a (universalisable) maxim about the (Kantian) rationality of helping the ship-wrecked. So far, this seems similar to our notion of solidarity-with-the-shipwrecked. However, a Kantian's universalisable logic means that she cannot pick and choose *between* maxims consistent with this logic. To give an example, if visiting cancer patients in hospital is a Kantian maxim, and so is donating to the Lifeboat Service, the Kantian is duty-bound to do both. Thus, one characteristic of solidarity (as perceived here) that sets it apart from Kantian duty is the former's contingency; the possibility that one can be disposed to visiting cancer-patients but not to donating to the Lifeboat Service, even if both are demanded by similarly universalisable maxims. This difference flows onto a second one.

When a Kantian visits a cancer patient, it is conceivable that she does so without love, pity, pleasure in helping a sick person, or from being in her company.

36

She visits because she must, in precisely the same manner that she is honest because of a maxim that prohibits lies. However, here lies a paradox. The patient is less likely to be helped by the Kantian's visit if she feels that it is performed coldly, out of duty, and without empathy. In Smith's (1759) words, a '... benefactor thinks himself but ill requited, if the person upon whom he has bestowed his good offices, repays them merely from a cold sense of duty, and without any affection to his person'. The Kantian knows this but is structurally unable to pretend to care personally (when she does not) because her visit is motivated by exactly the same

'force' that causes her to be honest, to respect red traffic lights and, of course, to visit cancer patients.

It might be argued that the same paradox emerges when someone visits our patient out of solidarity; motivated by empathy not toward her individually but due to her 'condition'. Not quite. Although solidarity is also impersonal in this sense, it differs crucially from Kantianism because the 'condition' responsible for it is not pre-determined by some steely, universalisable logic. The patient sees that her visitor is perfectly capable of disregarding all sorts of high-minded maxims (e.g. she lies when it suits, jumps red lights when impatient, ignores pleas for donations from the Lifeboat Service). And yet, her visitor is moved by the plight of cancer sufferers like herself. This inconsistency that solidarity allows for (and Kantianism bans) makes for a more fruitful hospital visit.

To recap, team-reasoning confines generosity to team members; natural sympathy limits it to those for whom we feel *as persons*; and Kantian generosity recognises no special entitlements to one's generosity. A Humean's $\lambda_i > 0$ can only be attributed to *i* thinking of the sufferers' ends as a *means* to *i*'s own; a Kantian's $\lambda_i > 0$ reflects *i*'s eagerness to treat *all* others as ends-in-themselves. And while the former will only be generous to *persons* whose interests she can adopt as her own, the Kantian ends up performing her 'duty' to all but lacks in real compassion. By contrast, the notion of solidarity steers a middle course. It identifies a *condition* which makes those who 'suffer' it worthy of one's generosity *independently of who they are and what interests they have*. Some misfortune beyond their control defines a group of *N* persons as those entitled to one's λ -generosity; thereafter, the agent feels an emotionally charged urge to help them *out of solidarity with their condition*. And because the selection of this *condition* does not derive from some rationally determinate formula, solidarity packs the emotional element that Kantian duty is missing.

DEFINITION 7.5 A person's σ -solidarity toward some group *N* is given as

$$\sigma_i = \begin{cases} \lambda_i & \text{iff conditions (I) to (IV) apply} \\ 0 & \text{otherwise} \end{cases}$$

(I) Personality-invariance: i selects target group N *independently* of any personal characteristics of its members

(II) Condition-specificity: Target group N is identified on the sole basis of an *adverse condition* which is *shared* by N 's members. This *condition* is selected by an unspecified, non-universalisable method

(III) Belief-irrelevance: λ_i is independent of beliefs $(\Lambda_{Nj}, \Lambda_{M \sim i}, \Xi_{M \sim i}, \Xi_{Nj})$

(IV) Non-instrumentality: Agent i 's choice of the set of persons N is irreducible to the maximisation of expected net gains from the future behaviour of others (N -members and non- N -members)

Condition (I) differentiates solidarity from utilitarian altruism, personal sympathy etc. by ruling out personal motives and interests as a possible source.

Condition (II) identifies solidarity exclusively with generosity directed at victims of misfortune, rather than of serendipity.

37

It also allows for a narrow and highly subjective focus of one's solidarity (by the virtue of the non-universalisability of the selection criteria) which is consistent with the often puzzling observation that a sighted person, who has no blind friends or relatives, may be prepared to go to incredible lengths to help with the education of blind children while remaining distant from similar efforts with deaf children. Condition (III) reflects the thought that solidarity cannot be motivated by an urge to impress others, or conform to their expectations (calculative or normative). Indeed it requires an autonomous moral judgment that some group N is somehow entitled to one's generosity, even if no well-recognised principle of justice so prescribes.

38

Condition (IV) is technically redundant (since a positive λ always comes at a personal cost – see prelude to *Definition 1*) but is included here in order firmly to remind us that we exclude from the realm of solidaristic acts those which, in the final analysis, are no more than shrewd self-interested investments.

So far we have established that, courtesy of our four conditions above, solidarity has been decisively distinguished from the previous section's other-regarding categories (b), (d), (e) and (f). The same conditions disqualify explanations (a) and (c) (team-reasoning and Kantian duty).

39

Conditions (I) and (II) ensure that σ -solidarity remains irreducible to team-reasoning since the λ -generosity underpinning it is not due to i belonging to target group N . Condition (II) keeps σ -solidarity analytically separate from some variant of Kantianism by introducing contingency into the selection of the 'condition' that motivates it. Taken at once, these conditions forge a notion of *solidarity* which can be juxtaposed usefully against the related ideas regarding fairness and justice. Such a juxtaposition, however, falls outside the scope of this paper.

40

The urgent question that needs to be addressed next derives from Condition (II). If not on a basis of a universalisable principle, how does one select the condition that motivates her solidarity? While different 'conditions' tussle for our 'targeted empathy' (e.g. 'shipwrecked', 'loser in a lottery', 'redundant worker', 'refugee', 'victim of torture' etc.), only a small number, if any, succeed in eliciting σ -solidarity. This eclecticism lends emotional and moral weight to the ensuing acts of λ -generosity (e.g. makes hospital-visiting worthwhile) but also calls for an explanation. Why are some moved by the plight of the deaf, others by the plight of the blind, while many more remain unmoved by either? This paper offers no definitive answer. [Perhaps there *can* be no such answer if Condition (II) is to be met (i.e. the selection process is not unique and

thus non-universalisable).] What it does claim, however, is (a) that σ -solidarity is probably as rare a phenomenon as it is socially important, and (b) that the reasons for selecting *the* condition(s) on which our solidarity trades may be either internal or external to our preferences.

Beginning with (a), there is little doubt that Conditions (I) to (IV) will remain dissatisfied more often than not. Most acts of generosity violate *personality-invariance* (in that they are directed to kin or friend); are *belief-contingent* (i.e. people are motivated to perform them because they are expected to); and verge

on the *instrumental* (e.g. sacrifices are seldom independent of the hope that it will be reciprocated). However, just as dishonest acts trade on the fact that not everyone is dishonest, generosity that is not motivated by solidarity finds fertile ground on which to grow only in social settings where σ -solidarity has not been eradicated completely.

41

'Other-regarding' deeds, which deep down are self-serving, must always remain parasitic on something resembling either our σ -solidarity or Kantian duty. Indeed, if perfectly egotistical acts can masquerade as other-regarding, selfless, solidaristic etc., this is so only because σ -solidarity not only makes sense but is also possible (and perhaps easier to relate to than Kantian high-mindedness).

Turning to (b), *i*'s choice of some 'misfortune' or 'adverse condition' with which to empathise can be motivated by two types of explanation. An internalist explanation is fundamentally Humean in that it places the burden of explanation on the evolving passions and the feedback effects between the latter and the corresponding social conventions (or 'equilibria') that they spawn. Of course, there are a variety of explanations consistent with this. For instance, a neo-Humean might argue that a rich tapestry of solidarity is woven gradually over time (e.g. some people develop solidaristic feelings toward the homeless, others toward the refugees etc.); its genesis resembling the spontaneous emergence of conventions in indeterminate social interactions while its survival depends on how successfully it regulates social life. In effect, neo-Humean solidarity (just like all other conventionally evolved patterns) adds to the evolutionary fitness of the community within which it sprung and, in a never-ending circle, is strengthened by it.

42

Of course, internalist accounts are not all neo-Humean. For instance, consider the following two-stage, rule-utilitarian account of *i*'s σ -solidarity toward members of group *N*: In the first stage *i* selects the condition which determines set *N* (e.g. those who are 'shipwrecked', 'HIV carriers' etc.) on the basis of some principle external to both her preferences and to any social expectations. In the second stage, *i* chooses $\lambda_i = \text{argmax}_\lambda (U_i[u_i(\lambda_i), W_N(\lambda_i)])$. Conceptually this two-stage process resembles Frankfurt's (1971) idea of a two-tier deliberation process for rational agents: one (the lower tier) where preferences determine outcomes and another (the higher tier) in which principles external to preferences decide which of the lower-tier deliberations should be 'trumped' and which should be allowed to pass.

By contrast, those arguing in favour of fully external reasons for action (see Hollis, 1987, 1998) might insist that genuine solidarity requires a moral psychology which enables *i* to distance herself completely from her own preferences and passions; to show her solidarity to *N*-members for reasons pertaining to *them*, rather than reasons appealing to some desire or urge in her own bosom. Most economists would dismiss this idea and would associate non-optimising choices with bounded rationality. This is due to their insistence that reasonableness reduces to CKIR or (in the term coined by Hollis, 1998) to philosophical egoism. However, there is no reason why this identification should be taken for granted. Unlike *homo economicus*, reasonable people can pass judgement

on their own passions or desires and one way in which they rebel against the

tyranny of preference is to do what is 'right' by some group of persons who are 'entitled' to their generosity. To the extent that this 'rebellion' is expressively (as opposed to instrumentally) rational,

43

and indeed finds expression in solidarity with sufferers of some misfortune, human motivation is under-explained unless solidarity is acknowledged as an important and distinct aspect of the human experience.

7.5 Radical solidarity: empathising with the victims of social power

In the previous sections solidarity was defined as empathy with persons afflicted by some shared misfortune (e.g. cancer victims or shipwrecked sailors). When the latter is a social artefact, as opposed to an accident of nature, solidarity turns radical. The nineteenth century anti-slavery movement, for instance, was an expression of radical political solidarity with the victims of humanity's darkest artifice. It is a general tendency of human societies in all places and at all times to generate social power structures which place whole groups of people, quite arbitrarily, into 'unfortunate' roles and situations. Spontaneously, and through no fault of their own, they become victims of an evolved social force which expels them to the periphery of social life. A disposition toward making sacrifices on their behalf will be defined below as *radical solidarity*. First, however, we need to define arbitrary, evolved, social power and the hierarchies which it fashions.

DEFINITION 7.6 Suppose that the distribution of resources and social roles within a community Γ is determined by a series of interactions between its members. Suppose further that Γ is subdivided in at least two groups *arbitrarily*; that is, according to criteria irreducible to differences in their personal talents, application, or 'worth'. Members of group $K \subseteq \Gamma$ are said to exercise two types of power over members of group $N \subseteq \Gamma$:

44

(a) **structural social power** if the structure of the interactions is consistently biased in their favour (and, therefore, so are their outcomes),

45

(b) **conventional social power** if the outcomes of interactions between agents $i \in K$ and $j \in N$ conform to some discriminatory *evolutionary equilibrium* even though the interactions are structurally symmetrical.

46

DEFINITION 7.7 *Radical* or *p-solidarity* is defined as σ -solidarity (see Definition 7.5

) directed *consciously* to those who live under the structural or conventional social power of others. More precisely, i 's [$i \in M \subseteq \Gamma$] radical solidarity ρ_i equals σ_i if and only if when directed to some group N for the reason that the latter's members are subjected to group K 's social power. Otherwise, $\rho_i = 0$ (even if $\rho_i > 0$)

Game 7.1 1-2-3. Pure strategy equilibria in bold

Strategies	left	middle	right
up	1, 3	0, 0	0, 2
middle	0, 0	2, 2	0, 0
down	3, 0	3, 0	3, 1

Game 7.2 Hawk–dove–cooperate. Pure strategy equilibria in bold

	h	d	c
h	-2, -2	2, 0	4, -1
d	0, 2	1, 1	0, 0
c	-1, 4	0, 0	3, 3

As an example, consider two games being played repeatedly among different identical opponents drawn from a large community Γ . Game 1-2-3 has a unique equilibrium (Nash and evolutionary) which awards payoffs 3 and 1 to the players selecting among the rows and among the columns respectively. A case of *structural social power* emerges if some social process systematically selects K -players to choose among the rows in meetings with N -players. By contrast, *hawk-dove-cooperate* (HDC hereafter) is symmetrical and features two equilibria in pure strategies [(2, 0) and (0, 2)] and one in mixed strategies (play h with probability $1/3$ and c with zero probability). Because of its symmetry, this game leaves room only for the *conventional* type of *social power*.

Although symmetrical in terms of its payoff structure, HDC spawns asymmetrical and highly discriminatory evolutionary equilibria *even if players are identical* in every respect other than their group membership.

47

As long as group membership is observable, it can be used as a behaviour-conditioning device whenever player $i \in K$ interacts with $j \in N$. To see why, consider the first few rounds during which differences in behaviour between the two groups can only be due to randomness. Once one of the two groups is *observed* to have selected h with higher probability (for reasons similar to why three tosses of a fair coin may yield three tails), a bandwagon effect begins to roll intensifying the originally random intergroup differences in aggression.

48

When the evolutionary equilibrium is reached, one of the two groups (say K) dominates the other (say N) in that its members play h consistently against members of the other group who acquiesce (i.e. respond with d). Thus, the latter are, by *Definition 6*, subject to the *conventional social power* of the former.

In 1-2-3 the process manufacturing the subservience of N -members works through the assignment of row-column social roles; an assignment which has not been explained here. One possible explanation of its origins can be based on a fully endogenous analysis in the context of a conflictual interaction such as *hawk-dove*.

49

In such games, as discussed above, some groups gain the upper hand for reasons that have nothing to do with their personal qualities (see notes 47, 52 and 53). Indeterminacy conspires with asymmetry in order to spawn some non-rational social hierarchy. Once *conventional social power* has been established, the discriminatory conventions which it produces spread from one interaction to another, perhaps by the force of analogy, and determine the allocation of social roles in a manner which favours the already dominant groups. For example, the group that dominates in *hawk-dove* ends up with the row-role in subsequent plays of game 1-2-3! Thus, *conventional social power* may oversee and spontaneously lead to the creation of *structural social power*. In the process, whole groups of people are arbitrarily assigned the lesser roles and, through no 'fault' of their own, are subjected to the misfortunes reserved for them by unconscious, supra-intentional social design.

Juxtaposed against such evolutionary accounts, a number of interesting issues flow from our definition of *p-solidarity* as solidarity with the victims of discriminatory social design. For example, instinctively, the notion of solidarity-with-an-oppressor seems strained. Interestingly, and encouragingly, this 'strain' shows up in our taxonomy of solidarity above. Consider a case in which a group of socially powerful agents is threatened with loss of power and thus privilege; e.g. the dissolution of a Mafia-type organisation or white rule in South Africa. To the extent that their loss of privilege, wealth and status can be thought of as a 'misfortune' afflicting them as a group, there is nothing in our original definition of a *solidarity profile* (see *Definition 2*) to rule out

solidarity as targeted empathy toward (or within) such groups. Indeed, it is even possible that such sentiments qualify as σ -solidarity, provided the conditions of *personality-invariance*, *condition-specificity*, *belief-irrelevance* and *non-instrumentality* hold (see *Definition 5*).

50

The anomaly is however revealed when we submit these cases to the test of radical solidarity; a test which they cannot but fail since radical solidarity is directed solely toward groups who fall on the short side of evolved, arbitrary social power.

So it seems that our last refinement of the solidarity definition (ρ -solidarity) drives a wedge between the sentiments underpinning the collusion between holders of arbitrary social power and those shoring up acts of sacrifice (on behalf) of its victims. Things get messier, however, in the presence of interpenetrating patterns of discrimination, where the same group may be, at once, the victims in one type of interaction and the perpetrators in another.

51

And if discriminatory patterns have a tendency to survive by dividing and multiplying,

52

then evidence of ρ -solidarity and coercive collusion, whose purpose is to maintain some form of discrimination, may be found within most groups.

A related issue concerns the connection between philanthropy and solidarity. Whether, and to what extent, the philanthropist's motives can be deemed solidaristic depends both on her reasons and cognition of the beneficiary's situation. In our account, the identification of a group as worthy of her concern and sacrifice is the first prerequisite. To qualify for σ -solidarity, her motives must be untainted by a concern for what others expect of her, or what there is 'in it' for her (a 'condition' also imposed by Christian and other religions). And to meet the criteria of ρ -solidarity she must be conscious of the specific social design which manufactures

and arbitrarily assigns misfortune to undeserved victims. By these criteria, few Victorian philanthropists' acts and motives would qualify as *solidarity*

53

and even fewer for *radical solidarity*.

54

Perhaps the natural limit of *radical solidarity* is a capacity to focus one's endeavours on undoing the root-causes of others' systematic disadvantage and misfortune, even if this means undoing also the sources of one's own privileges. Such radical solidarity transcends mere palliative efforts; it threatens to dismantle whole networks of privilege and destitution but carries enormous risks for both 'donor' and 'recipient' as it combines opportunities for progress with the risk of gigantic folly characteristic of all radical change.

7.6 Epilogue

Hurley (1989) castigates *homo economicus* for lacking the *nous* effectively to engage in the bewildering enterprise of acting in a manner *organically* consistent with the objectives of the team to which she belongs. This chapter took Hurley's theme further by focusing on organic connections of the self with groups of 'others' to which one does *not* belong, linking the discussion with the book's broad theme on indeterminacy.

A rational person may expect nothing *of* a group of 'others' to whom she does not belong. She may care not one iota *for* them individually. She may feel no duty *to* them in particular. She may even detest the idea of belonging to their 'team'. And yet this person may, perfectly rationally, sacrifice a great deal for them. Clearly, this is a notion that neoclassical economics, to its detriment, cannot even begin to wrap its collective mind around. Indeed, the neoclassical economists' 'ideal man', *homo economicus*, only acts when there is something 'in it' for him, and would not lift a finger

on behalf of such a group of 'others' under the circumstances.

However typical of men and women the neoclassical humanoid may be, the very possibility of rational solidarity with 'others' bears an importance inversely related with the frequency of genuinely solidaristic acts. *Some* intelligent people, *some* of the time, are capable of selfless sacrifice, moved neither by expected gain nor altruism nor duty, but by a fierce repugnance for the suffering caused by some accident of nature or of social evolution.

55

And this matters to everyone else in society.

Of course, empirical observation cannot help us distinguish genuine solidarity from impostors, just as it cannot settle disputes between, say, Humeans and Kantians. Yet this does not lessen the importance of exploring philosophically the notion of authentic solidarity. For its very possibility, however faint it might be, provides the foothold necessary for shallower forms of solidarity to proliferate. Tiny as these ripples of genuine solidarity may be, they often turn into torrents of targeted empathy through imitation, social influence – even sheer hypocrisy. When they do, the social scenery is transformed and the cement of society is inserted between the bricks of individualist endeavours.

Neoclassical economic theory is a powerful tool for modelling behaviour in response to preferences inhabiting the well-defined space within the walls separating one self from an 'other'. Solidarity, on the other hand, refers to a phenomenon made possible because these walls are more porous than neoclassicism would permit; it alludes to a series of human interactions unfolding in the space *between these walls*, in a kind of no man's land where the plight of others inspires us to experiment with violations of our current 'preferences,' rationally toy with alternatives to the prevailing constraints of 'rationality,' throw away the mask of self-sufficiency, reach out for one another, re-discover something 'real' and authentic about our nature and, at rare moments, believe that there is more to us than some weighted sum of desires. Those of a romantic disposition may even conclude that solidarity-with-others is a prerequisite for throwing out a bridge over to our 'better' self; a capacity that neoclassical models of human agency must bleach out of humans before their practitioners can model human actions as the dehumanised 'moves' of stimulus-activated automata.

Central to the notion of solidarity developed in this chapter is the indeterminate choice of the group of 'others' that becomes the focal point of one's solidarity. By contrast, Kantian imperatives and neoclassical utilitarianism allow for 'other'-regarding acts that are *pre-determined* by the postulated model of men and women. The Kantian agent is constitutionally compelled to perform 'other'-regarding duties whenever the categorical imperative kicks in. Without any discretionary power of her own, for reasons that are external to herself, without even the need to harbour sentiments, e.g. empathy, for the recipients of her kindness. *Homo economicus*, equally, is compelled to act non-selfishly *if it is in his broader self-interest to do so*. But to experience authentic solidarity, at least as defined in the preceding pages, the agent must choose from among the groups competing for her empathy, the group or condition that she will direct her empathetic generosity toward. Moreover, and this is crucial, her choice cannot be predetermined, i.e. model-able, in any way. In this sense, indeterminacy lies at the heart of authentic solidarity, just as it underpins good theatre, art and music.

Notes

1

This chapter is based on Arnsperger and Varoufakis (2003).

2

There is hardly a European politician who, in the aftermath of monetary union, did not call for the blending of stringent monetary policies with a new commitment to solidarity with weaker members of society. Such calls were reinforced from an array of institutions ranging from the churches and social activist networks to the Confederation of European Industries (see Rouille d'Orfeuille, 2002). However, once the eurozone crisis began, solidarity turned into a

catchphrase for the bailouts of the banks of the surplus countries, which were 'marketed' under a deceptive cloak of solidaristic rhetoric regarding the importance of 'showing solidarity to the people of Greece, Ireland etc.'

- 3 Evolutionary biologists tell us that altruism is not a puzzle, in the sense that there is plenty of evidence from the animal world supporting the idea that altruistic behaviour does indeed improve a species' fitness (see Dawkins, 1976; Midgley, 1994). Economists favour models of enlightened selfishness in which bargain-hunting agents, though incapable of resisting the lure of a marginally higher payoff, are nevertheless led to the conclusion that it pays to be 'good'. Whilst this is the rational choice theorist's favourite explanation of humanity's mysterious, other-regarding side, it is by no means the only one. Some (see Sugden, 1986) still rely on Hume's (1888) distinction between selfish and self-interested actions, and the notion of conventionally reinforced natural sympathy that is founded on this distinction. Others turn to bounded rationality and evolved social reciprocity, as opposed to instrumental or economic reciprocity; that is, to norms of cooperative or seemingly altruistic behaviour which jump from game to game through analogy and habit (see Hoffman, McCabe and Smith, 1996). Non-utilitarian thinkers, meanwhile, have been focusing on explanations turning on kin selection, rationally deduced obligations to others (or duties, e.g. Kant, 1949) and ideas about justice and fairness (see Rawls, 1971).
- 4 'To breed an animal capable of promising – isn't that just the paradoxical task which Nature has set herself with mankind, the peculiar problem of mankind?' (Nietzsche, 1956).
- 5 Under the assumption of cardinal utilities, a particular case would be a Benthamite aggregation such that set N comprises the *complete* human population (and W_N is the average cardinal utility). Another particular case would be for set N to contain a single person: the one with the lowest utility (a type of welfarist-Rawlsian solidarity).
- 6 Liberals should beware the assumption that an act is 'generous' when the actor deems that she has benefited others through her own sacrifice. Sen (1970) issues an early warning. In our context it takes the form of a query: What if i feels that group N members need to be 'saved' from themselves by, for example, being burnt at the stake? Is burning them an act of kindness? A simple retort is that, naturally, it is anything but an act of kindness. But, on the other hand, if i genuinely thinks that she is benefiting them, we should accept that she is performing an act which she perceives, misguidedly of course, as kind.
- 7 Act a_i is generous ($\lambda > 0$) when both $s(a_i) > 0$ and $w(a_i) > 0$. When $s(a_i) < 0$ and $w(a_i) < 0$, we have an act that causes hurt at no expense to the agent and, therefore, $\lambda = 0$ even though $s(a_i) \times w(a_i) > 0$. Spiteful acts set $s(a_i) \times w(a_i) < 0$ as they imply $s(a_i) > 0$ and $w(a_i) < 0$. Product $s(a_i) \times w(a_i)$ is also negative in cases of reciprocal kindness occurs' i.e. when agent i benefits others [$w(a_i) > 0$] but does so expecting something back in return [i.e. $s(a_i) < 0$]. In both these cases (spite and reciprocity) *Definition 1* sets λ -generosity equal to zero. Finally, note that the intersection of groups N and M may well be non-empty.
- 8 For example, Rabin (1993) argues convincingly that the same action can be deemed fair or unfair depending on the agent's first- and second-order beliefs. Chapman (1998) takes this idea further by examining how rational behaviour might be affected if agents had to give well argued reasons for their actions; as they must in a court of law.
- 9 Geanakoplos *et al.* (1989), Rabin (1993) and Sugden (2000) model instrumentally rational actions which transcend the Humean divide which keeps beliefs separate from motives (e.g. utility). The common thread running through these three articles is that a person's valuation of a certain outcome depends, among other things, on her second-order beliefs (that is, on what she thinks her opponents/friends expect her to do).
- 10 Calculative or positive beliefs are mere predictions. We use these epithets in order to distinguish them from normative beliefs which pertain to beliefs regarding what *ought* to happen; as opposed to what *might* happen.
- 11 Note that this second-order belief is not a truly normative one. A truly normative second-order belief would correspond to what i thinks that j ought to think that i will do.
- 12 This game is identical in structure to Rousseau's stag-hunt game. Rousseau's original narrative had a group of hunters choosing between combining their efforts to catch a stag (the grand prize capable of feeding the group for days) or, alternatively, hunting skinny hares individually. The stag would escape if even a single hunter broke the 'chain' and sought to capture hares (i.e. everyone's payoffs is determined by the effort expended by the least committed members). Rousseau's point was that were the hunters to trust one another to pursue the stag diligently, they would all do so. However, pessimism about the group's solidarity would force them all to the suboptimal pursuit of hares. In recent times, experimental work has shown coordination to converge on inefficient outcomes in this type of game. It seems that Pareto-dominated Nash equilibria are selected because risk-dominance overpowers Pareto-dominance. See van Huyck, Battalio and Beil (1990).
- 13 Note that, unlike the prisoner's dilemma or the free riding game, there are no built in incentives in this game to cheat/defect. If one expects everyone else to contribute maximally one would follow suit.
- 14 Suppose the expected minimum choice equals m , but player i is prepared to choose $a_i = m + x$. The sacrifice involved equals x since sacrifice level $s_i = (A - 1)m - [(A - 1)m - x]$. When commonly anticipated, this sacrifice will lead all to make it. In this sense, i 's sacrifice x has increased the welfare of the rest of the group to the tune of $w = (N - 1)(A - 1)x$. Thus, i 's λ -generosity equals $\lambda_i = s_i \times w = x^2(N - 1)(A - 1)$. Under *minimal generosity*, the sacrifice is minimal, i.e. equals \square , and therefore $\lambda_i = s_i \times w = \varepsilon^2(N - 1)(A - 1)$; a value lower order, viz. the degree

of sacrifice involved. On the other hand, for λ -generosity to be of ε -order, $[i = \varepsilon = s_i \times w = x^2(N-1)(A-1)]$,
in which case the relevant sacrifice level is $x = \sqrt{\frac{\varepsilon}{(N-1)(A-1)}} + o(x)$.

15

This being a one-shot game, the 'algorithm' described here unfolds in logical, rather than in historical, time. It simply captures the train of thinking that leads players to the unique equilibrium (in a manner analytically identical to the process of iterative dominance or, as it is sometimes known, the successive elimination of dominated strategies).

16

Note that the difference between this variant of the game and the original is that here the average choice of number in the group has replaced the minimum choice in each player's utility function. Obviously this changes the character of the game from that of a coordination/stag hunt type to a N -person free-rider problem since, by choosing a number smaller than the average choice, your payoff rises as long as $N > A$. To see this, note that the derivative of player i 's payoff function u_i s.t. a_i is negative as long as $N > A$. And since there can be no fewer than 1 player, $N > A > 1$ is the condition under which each of the N players has a dominant strategy: 'Set $a_i = 1$!' In short, it pays to undercut the 'contribution' of the average player in the group.

17

Note however that the amount of generosity required to sustain the cooperative outcome varies. For if they all expect maximal generosity of each other, then the actual sacrifice of each $i \in M$ (s_i), and the welfare benefit of others (w) following this sacrifice, is smaller than it would have been if cooperation was not envisaged.

18

It is easy to see that cooperative behaviour requires $a = 10$, a value that maximises λ . Taking the limit as N tends to infinity, we note that, in games involving many players, a cooperative outcome requires mutual \square -generosity equal to $81A$.

19

For a summary of why instrumentally rational agents cannot be reasonably expected to choose a cooperative disposition in free rider (or prisoner's dilemma) interactions, see Hargreaves-Heap and Varoufakis (1995),

[Chapter 5](#)

20

For a modern version, complete with empirical evidence, see Andreoni (1990).

21

'If he was to lose his little finger tomorrow, he would not sleep tonight. But provided he never saw them, he will snore with the most profound security over the ruin of a hundred million of his brethren, and the destruction of that immense multitude seems plainly an object less interesting to him, than this paltry misfortune of his own.' Smith (1759).

22

Note that the passage from Humean to *homo economicus* is not as straightforward as some seem to think. Indeed 'sanitising' the passions so as to turn them into preferences (cardinal or ordinal) is philosophically problematic. See, for instance, Sugden and Hollis (1993). For a different perspective on the same issue, see Margolis (1981).

23

Nevertheless, the paradox of 'rational saints' remains. If each player is motivated by a selfless urge to satisfy the preferences of others, then in the context of a prisoner's dilemma agents may still get caught up in a mutual-minimum since each will be failing to make a sufficiently satisfying sacrifice on others' behalf.

24

An anonymous referee made the point that '... Kant meant us to ask ourselves whether our action is possible as such if all selected that action. Hence the categorical interdiction of lying and cheating, as one literally cannot cheat if nobody honours agreements...' This is not the place to enter into hermeneutical debates around what Kant really meant. However, it is fascinating to note that, if we were to accept the referee's interpretation,

[Proposition 7.1](#)

would be threatened. The latter shows that *minimal generosity* leads to an equilibrium in which generosity is rendered impossible. In a sense, Kant would be censoring not only lies but also contributions to the Public Good.

25

We say 'minimum' because there is nothing stopping a Kantian from boosting her generosity beyond the level determined by her 'duty' in cases in which she does feel sympathy for the target group or person.

26

Rawls' (1971) argument is that rational agents will exercise infinite risk aversion behind the veil and will thus choose the best outcome from the perspective of the person who will end up being worst off. Thus, if agents are forced to go behind the veil, and choose while there, their choices (which amount to a maximal λ) are deemed, by Rawls, to be merely rational. However, in view of the fact that no one is ever forced to go behind the veil, a willingness to decide what to do on the basis of what one would have done *had one found oneself behind the veil*, is a willingness tantamount to a *generous* predisposition.

27

Akerlof (1980) utilises this idea in order to model the decision of unemployed workers not to undercut the wages of their employed colleagues and Varoufakis (1989, 1990a) tells a story about wage and employment determination when a trades union's power stems from worker solidarity during (actual or threatened) strikes.

28

Geanakoplos *et al.* (1989) examine a situation in which person A must choose between acting courageously or cowardly (NB this is not really a game in the sense that there is only one player: A). Her utility from these two outcomes hinges crucially on what others' expect of her. So, if A believes that others expect her to act courageously,

she will *want* to do so. If not, she will prefer to act like a coward. There is nothing to suggest that in the former case A's utility will not be lower than in the latter.

29

Rabin (1993) labels a similar situation an *un-fairness equilibrium*.

30

For example, suppose that for $i \in N$ the utility function is given by: $U^i = u_i(\pi_i, \lambda_i) + \gamma_i[\lambda_i \times \Lambda_{Mj}]$ where π_i is i 's material payoff and $\gamma_i > 0$ is some constant which reflects i 's relative valuation of the means by which certain payoffs are produced. Similarly, let $U^j = u_j(\pi_j, \lambda_j) + \gamma_j[\lambda_j \times \Lambda_{Mi}]$ be the utility payoffs to $j \in M$. Such a maximand instructs i and j (as long as the γ 's are large enough) to set $\lambda_i, \lambda_j > 0$ if they anticipate $\Lambda_{Mj} > 0$ and $\Lambda_{Mi} > 0$ respectively. However it also urges them to set their $s > 0$ in order to cause $w < 0$ (i.e. to make positive sacrifices in a bid to hurt the other group) if they expect a similar disposition from members of the other groups.

31

See also Sugden (1982).

32

Three subjects A, B and C participated in a lottery which would award each DM10 with probability 2/3. Subjects were asked ex ante to state how much of their winnings they were prepared to share with the other subjects in their team of three who won nothing. Subject A was invited to declare the sum she would donate to B (or C) if A were to win DM10 and B (or C) was the only loser in the trio. Let us call this sum X . Then A was asked to select her donation to both B and C if neither B nor C were to win any money. Let this sum equal Y and assume that 'losers' B and C split Y between them. 52 per cent of the subjects chose $X \leq Y$ (up to rounding error), a finding which the authors label fixed total sacrifice (FTS) and show to be inconsistent with standard utilitarian altruism.

33

For instance, in the Selten and Ockenfels experiment, symmetry means that, in A's eyes, *ceteris paribus* the loss of one expected currency unit (e.g. DM1) by a 'losing' subject B yields the same disutility for subject A as the loss of DM1 by a winning C who nevertheless donates DM1 to some other 'loser'.

34

The willingness to make a sacrifice on behalf of others based on the expectation that, if roles are reversed, members of this target group will/should come to one's aid.

35

By 'enlightened selfishness' we mean generosity motivated by the (selfish) hope that the beneficiary will re-pay the donor in the future. Furthermore, utilitarian altruism requires a specific person's utility to be introduced as a variable in the donor's utility function. But our definition of solidarity rules out person-specific motivation in two ways: First, by identifying solidarity as a subset of λ -generosity (which in itself rules out self-serving sacrifices as potentially λ -generous acts); Secondly, by tying solidarity up with other peoples' condition, rather than with their disutility from it.

36

There is of course no doubt that a Kantian motivation may coincide with feelings of love, sympathy etc. However, Kant's point is that even when the latter are absent, the visit ought to take place. Our interest lies in the effects and nature of such purely Kantian acts of generosity.

37

For example, i might be λ -generous to a group of pop-stars that she worships. However, given condition (I) this does not qualify as a case of σ -solidarity.

38

Thus, norm or custom-following [à la Akerlof (1980) and Varoufakis (1989)] do not qualify as examples of σ -solidarity. In this sense nor do the concerns for one's image within a group mentioned by Olson (1965) or Becker (1974) since, according to our definition, σ -solidarity is irreducible to social norms or public expectations.

39

Effectively, we argue that, whenever $\lambda_i > 0$ but $\sigma_i = 0$, the explanation of i 's λ -generosity must be sought in some of the other-regarding categories in Section 3.

40

We believe, nevertheless, that σ -solidarity has important implications for justice: According to one perspective on justice, the latter flourishes when altruism reaches its limits. It comprises a set of constraints regarding our behaviour toward persons for whom we harbour no natural sympathy (for if we did, we would not need moral constraints in our dealings with them). In this paper we argue that something else is also born, in addition to justice, at the limits of altruism: Solidarity! It pertains to instances of sacrifice and generosity motivated by 'worthy causes', rather than by an altruistic urge to contribute to specific individuals. The single mother of our Boat Service example may feel no ethical obligation to yachtsmen on the grounds of any principles of 'justice'; and yet, she may contribute in response to an antipathy toward the abstract idea of a lone figure helplessly fighting a losing struggle against menacing seas. Similarly with the subjects in the Selten and Ockenfels (1998) experiment: Solidarity with the losers is a feeling quite distinct from a commitment to fairness. The interaction between solidarity and justice is an obvious area of further study.

41

Sugden (1993) describes instrumental accounts of moral behaviour as: 'parasitic on moral theories that enjoin us to behave in ways that are not instrumentally rational' (Sugden, 1993). Thus, the presence of even a small percentage of persons capable of σ -solidarity may be the necessary initial condition for some bandwagon to start rolling (e.g. Akerlof, 1980, or Varoufakis, 1989).

42

Though not narrated in terms of solidarity, Sugden's (1986) main thesis is consistent with this account.

43

For a discussion of expressive, versus instrumental, rationality see Hargreaves-Heap (1989).

44

Hereafter the analysis will proceed on the assumption that the two groups do not overlap. However, the analysis generalises naturally when there are more than two groups and a person can belong to more than one at the same time.

45

For example, a game with a unique equilibrium which awards higher payoffs to K -players than to N -players.

46

For example, a symmetrical game with twin equilibria one of which favours the K -players, the other the N -players. If a convention evolves selecting the former equilibrium, K -players will, according to Definition 6, enjoy conventional social power over N -players. And vice versa.

47

For the theoretical proof see Weibull (1989). Hargreaves-Heap and Varoufakis (2002) report on an experiment which confirms this theoretical intuition. In it, players were divided in two groups ('red' and 'blue') and only their colour was made known to their opponent. And yet, in repeated play of the hawk-dove game, one of the two groups (in some sessions the 'red,' in others the 'blue') emerged as dominant. When later they played the HDC game above, the same pattern continued with one important difference: when dominant colour players were matched with one another, they never cooperated whereas when disadvantaged colour players met, they cooperated most of the time (a case of solidarity among the discriminated?).

48

Selecting h can be interpreted as aggressive behaviour, d as acquiescent and c as cooperative.

49

For example, in the context of conflict over property rights.

50

There is, for instance, plenty of documented evidence of selfless, reciprocal sacrifice among the ranks of otherwise abhorrent groups and organisations (e.g. SS officers).

51

Much ink has been expended in an attempt to come to terms with situations in which, for instance, the male victims of racial discrimination struggle to retain their exercise of arbitrary social power over their wives, mothers and sisters. In the sense of this paper, they pose simultaneously as the potential recipients of π -solidarity (in interactions with the white community, labour market etc.) and as parties to a collusion which fails the conditions of π -solidarity outright.

52

See Hargreaves-Heap and Varoufakis (1995),

[Chapter 7](#)

, for an evolutionary model of how discriminatory conventions gain evolutionary fitness through division and multiplication.

53

Since most philanthropical activity was part of the facade of Victorian socialising.

54

Since the last thing on most Victorians' minds was the social process manufacturing systematic, large-scale deprivation. Instead, they tended to focus on the personal responsibility of the wretched and the poor for the condition they found themselves in.

55

Of course, an economist might argue that the amelioration of the repugnant suffering, and the indirect utility so procured, is the solidaristic agent's reward. This is neither here nor there. Whether the reason for acting in solidarity with an 'other' is internal (e.g. indirect utility) or external to one's preferences is too rarified a question to delve into here.

8 On the power of what others think

How indeterminacy explodes when our preferences are influenced directly by other people's beliefs

8.1 Prologue

8.1.1 Background briefing

It takes a total indifference to the human condition not to notice that there is something amiss with the model of men and women at the heart of neoclassical economic analysis. However, students give their textbooks the benefit of the doubt, assuming that more sophisticated versions of this model await them once they grasp the intricacies of the simpler version before them. That was, indeed, my own hope when I first came across *homo economicus*; in short, I imagined that, at the cost of some extra mathematical complexity, he could be ... civilised.

Chapter 2

, the reader may recall, was my first attempt in this direction. In a bid to capture a worker's psychological costs of breaking an industrial strike, I employed a utility function that contained not only material rewards (such as the wage and the likelihood of being fired) but also a psychological component that turned negative when one crossed picket lines (and which was inversely related to the proportion of one's colleagues that also broke the strike). The results of that theoretical endeavour were presented in subsequent chapters and I shall say nothing more about them here.

The present chapter takes the ambition to improve upon *homo economicus*' psychology onto a higher level. So far the psychological utility that we have allowed into the model was action-based or, equivalently, outcome-based. For example, a trades unionist's psychological utility from not crossing a picket line was determined fully by: (a) the fact that she did not cross the picket line herself and (b) the proportion of her colleagues that did. Of course even this rather base form of psychological utility sufficed to throw the model into the clasps of radical indeterminacy. Imagine what happens when the psychological make up of our agent is allowed to become richer, more sophisticated...

In the following sections a different species of psychological effect is allowed: one's evaluation of a certain outcome now depends not just on how she acted (in association with how the other participants acted) but is also influenced, at least partly, by others' expectations of her. This subtle but crucial complication of the interaction turns on the juxtaposition between (a) caring about what others think exclusively because you know that what they think informs how they will act

(which is the only thing you really care about), and (b) caring about what others expect of you directly; because, for example, you do not want to disappoint them, or you do not want to conform to their expectations. In the more 'primitive' case, (a), others' beliefs matter because knowing them helps you predict what they will do. In the second case, (b), others' predictions of your behaviour affect directly what you want the outcome to be!

In the rest of the chapter we shall examine how a direct link between second-order beliefs and preferences changes game theory, making it far more interesting but also causing indeterminacy to reach a new, hitherto unknown, crescendo; a degree that scares the living daylights out of the average neoclassical economist. The more sophisticated neoclassicist, meanwhile, performs the *dance of the meta-axioms* majestically so as to 'close' the model by ushering in the most formidable, ironclad version of the third meta-axiom. But more on this in the chapter's epilogue.

8.1.2 The rest of the chapter

It is tempting to think that re-working game theory's model of individual agency might

lead to a more wholesome theory of social agency. How will game theoretical results be altered once we move to a richer human ontology (i.e. a better model of the human person)? Will the *prisoner's dilemma* continue to resist cooperation as doggedly as it has in standard game theory? Will mutual acquiescence remain 'irrational' in *hawk-dove*-like interactions?

In this chapter we investigate some interesting attempts to 'complicate' the players' psychology and, in particular, their motivation. Let us begin with the one-shot *prisoner's dilemma*, where strategy *d* is to 'defect' and *c* to 'cooperate'.

Game 8.1 The prisoner's dilemma.

	<i>d</i>	<i>c</i>
<i>d</i>	1, 1	4, 0
<i>c</i>	0, 4	3, 3

Suppose that the payoffs are dollars. Does this mean that rational players must defect (play *d*)? If the payoffs were utils, the answer would be affirmative. For the assumption of game theory is that utils motivate exclusively. However, dollars do not motivate exclusively. For instance, a player may 'like' dollars but she may also like something else – say, 'fairness' or 'equity'. Game theory has no quarrel with this thought. Indeed, it insists that, before we study the game's strategic structure, we ought to convert dollars into utils.

Here is one example of the possible psychological payoffs associated with equity. Suppose that player *i* values dollars (\$) but that she also dislikes unequal distributions of dollars between herself and her fellow player (*j*). Then her utility looks like this: $U_i = a\$_i - b|\$_i - \$_j|$, with ratio b/a reflecting *i*'s valuation

of equity (or fairness) relative to own dollars. Suppose now that two players who share these preferences meet in the context of

Game 8.1

. Translating the dollar payoffs into utilities yields

Game 8.2

– a totally new game:

Game 8.2 The prisoner's dilemma when players value equity.

	<i>d</i>	<i>c</i>
<i>d</i>	a, a	$4(a - b), -4b$
<i>c</i>	$-4b, 4(a - b)$	$3a, 3a$

If the players' dissatisfaction from receiving a dollar more or less than her opponent exceeds (in absolute terms) her satisfaction from gaining 25 cents (i.e. $b/a > 1/4$), then the game ceases to be a *prisoner's dilemma*. To be precise it becomes a *stag-hunt* type of game featuring two Nash equilibria (*cc* and *dd*).

What has happened here is that the players' psychological, ideological, or moral utility from equity altered the game's strategic structure, rendering *c* a best reply to *c* (as long as $b/a > 1/4$).

Homo economicus is, therefore, unperturbed by accusations of having no psychology and no capacity to overcome prisoners' dilemmas cooperatively. What we can say, however, is that his psychological preferences are fixed before the fun and games begin; that his psychology is independent of what others think (and what he thinks that they think). He may have preferences that are ethical, moral or even psychological. However, he has no preferences that depend on what is expected of him

to think or do.

In summary, this chapter will focus on the analytical effects of making players psychologically more complex, while retaining the assumption that they are instrumentally rational. Will they still behave as game theorists expect them to? Or will alternative behavioural patterns emerge? Moreover, will this complication of *homo economicus*' psyche narrow down the range of potential equilibria, or will the problem with *indeterminacy* spin, even more hopelessly, out of control?

The conversion of dollars into utility, which transformed the *prisoner's dilemma* ([Game 8.1](#)

) into

[Game 8.2](#)

, can be generalised mathematically, for all sorts of games, in the following simple manner: Utility is given as the sum of two sub-utility functions: $M(\cdot)$ and $\Psi(\cdot)$,

$$U_i(O) = M(O) + \Psi(O, B^2) \quad (8.1)$$

where $M(O)$ denotes the utility (or degree of preference-satisfaction) resulting *directly* from player i 's *material gains* from outcome O , and $\Psi(\cdot)$ denotes the *psychological utility* from the same outcome. What makes the following analysis interesting, is that $\Psi(\cdot)$ will no longer depend solely on outcomes, O , but also on B

², i.e. the player's second-order beliefs, which are her prior beliefs about what others expected her to do.

[Section 8.2](#)

explores some forms that $\Psi(\cdot)$ might take so as to capture the power (on us) of what (we think) others expect (of us). Both others' predictive and normative beliefs are looked at, as a source of motivation, before turning to Rabin's (1993) model of what we think determines beliefs of what one is 'entitled' to when in a particular role of some game. These perceived entitlements are important because they colour our views of what we, and others, have a 'right' to expect in any strategic interaction, a set of views that is crucial in determining the power of beliefs to shape our preference ordering of the various feasible outcomes. (

[Section 8.2.5](#)

presents my own formulation of these entitlements.)

[Section 8.3](#)

combines evolutionary theory with psychological game theory in an attempt to tell a story of how our ideological-cum-psychological beliefs evolve; of how our perception of each player's entitlements shift and change in response to evolving behavioural patterns and norms. Lastly,

[Section 8.4](#)

concludes the chapter and ties up its findings with the book's overarching narrative on indeterminacy.

8.2 The motivating power of second-order beliefs

8.2.1 Burdened by others' predictions

Let us start from a definition of second-order predictive beliefs:

Second-order (predictive) beliefs: Suppose player A chooses between strategies (s_1, s_2, \dots, s_m) with probabilities (p_1, p_2, \dots, p_m). Suppose further that, before observing A's actual choice, B's estimates of probabilities (p_1, p_2, \dots, p_m) are given as (p'_1, p'_2, \dots, p'_m). We define A's second-order beliefs (q_1, q_2, \dots, q_m) as her estimates of (p'_1, p'_2, \dots, p'_m). For example, if $p'_i = 1$ and $q_i = 1/2$ this means that B predicts that A will choose strategy s_i with certainty but A wrongly thinks that B expects her to do so (i.e. choose strategy s_i) only with a 50 per cent probability.

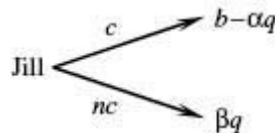
Consider now

[Game 8.3](#)

below. Jill must choose between strategy c and nc and assume that she chooses nc with probability $p = \Pr(nc)$. Meanwhile Jack is observing Jill and, before her choice is revealed, predicts that she will choose c with probability $p' = E^{\text{Jack}}(p)$. Jill knows Jack is 'watching' and cares deeply about his expectation. Indeed, if she thinks that he is expecting her to avoid c (i.e. Jill predicts that p' is high), she suffers psychological disutility from frustrating his expectations by playing c . Let $q = E^{\text{Jill}}(p')$ be Jill's estimate of the probability with which Jack expects her to shun c . In terms of the arguments in utility function (

8.1

) we can capture this situations simply by having $M(c) = b$, $M(nc) = 0$, $\Psi(c) = -\alpha q$, $\Psi(nc) = \beta q$. Putting the whole utility representation together, we end up with the following 'game':



Game 8.3 Jill's dilemma

where α is a positive constant capturing the rate at which Jill's psychological utility will decline if she chooses c as her estimate of Jack's prediction that she would not choose c increases; and β is another positive constant reflecting her psychological utility gains when she desists from c the more confident she is that Jack is not expecting her to adopt c .

An example: Let c be some act that Jack considers corrupt or 'immoral' and, also, one that Jill knows that Jack thinks ill of. If q is close to one, Jill thinks that Jack is expecting her to refrain from the shady deed. In this case, if she goes ahead with it, she will collect the material payoff (b) but will lose psychological utility in proportion to q . She gets the material rewards ($M > 0$) but feels bad at having frustrated Jack's prediction that she would prove 'upstanding' ($\Psi = -\alpha q$). On the other hand, if she turns her back on the ill-thought practice (c), and chooses nc , she will lose the opportunity for material enrichment ($M = 0$) but will feel better from psychological utility reward $\Psi = \beta q$.

Clearly, as q falls below a certain threshold (q^*), Jill's balance of utility gains and losses is tipped in favour of behaviour c . The greater Jill's belief that Jack expects her to avoid c the less her relative valuation of the material payoffs from c .² Note that Jill is the only active player in this 'game,' since Jack enters the fray only indirectly and through Jill's expectations of what he predicts that she will do. Nonetheless, the fact that the decision maker's payoffs are determined directly by her (second-order) beliefs makes it possible (indeed necessary) to demarcate outcomes that are consistent with beliefs. This notion of a *psychological equilibrium* was devised by Geanakoplos *et al.* (1989). In the case of

Game 8.3

it is easy to show that there are three such psychological equilibria:

The three psychological equilibria of

Game 8.3

- (i) $p = p' = q = 1$ Jill chooses nc and collects only psychological payoff β
- (ii) $p = p' = q = 0$ Jill chooses c and collects only material payoff b
- (iii) $p = p' = q^*$ Jill randomises and collects on average payoffs equal to:

$$\frac{b}{(\alpha + \beta)} \left[\frac{\beta b}{(\alpha + \beta)} + [1 - \{b/(\alpha + \beta)\}] [\beta - \{\alpha b/(\alpha + \beta)\}] \right]$$

Remarkably, we ended up with multiple equilibria in a game where only one person makes a single move! Any of these three situations is consistent with Jill's instrumental rationality and the requirement that her beliefs are consistently aligned with Jack's. One conclusion is that indeterminacy managed to spread to a single player game courtesy of that player's direct concern for what some observer thinks.

A second conclusion comes forward the moment we delve into Jill's preferences over these three equilibria. It is instructive to consider the case in which the fact that b exceeds β gives Jill a reason to prefer equilibrium (ii). This preference means that, if Jill had a choice (which she does not), she would rather that Jack expected her to opt for c (as opposed to expecting her to choose nc). Jill would then play c , liberated from the heavy psychological losses ($-\alpha$) that would arise

from choosing c against the grain of Jack's expectations. But, she cannot control Jack's thoughts. If q is high, Jill will be caught in a never-ending circle of expectations that she cannot short-circuit and which keep her from choosing c . Jack's expectations that she will shun c are confirmed only because this is what he expects. For had he expected c , Jill would have been all too glad to oblige.

8.2.2 Labouring under others' normative beliefs

So far only *calculative*, or *predictive*, second-order beliefs have entered directly into the agent's motivation. Jill's will was affected by what she thought Jack was *predicting* of her. But what if the motivating beliefs are of the normative variety? Before the battle of Trafalgar, Lord Nelson famously told his men that he 'expected' great deeds of them. Rather than mere prediction, his statement constituted a morally charged incitement.

In principle there is no analytical problem in converting the previous section's model so that the beliefs that enter into Jill's utility are of the normative kind. Nothing stops us from thinking of probability q as Jill's estimate of Jack's belief on what Jill *ought* to do (as opposed to what she *will* do). Rabin (1993) takes things much further by postulating that, not only are we affected by others' predictions, but that we are swayed by our interpretation of others' *motives* as well.

Rabin's main idea is that, while remaining fully instrumentally rational (something which we continue to know 'commonly'; i.e. CKR is still imposed) our *psychological utility from a certain outcome* (that is, from a particular combination of strategies) *depends on our beliefs about our opponents' motivation*. Consider again the *prisoner's dilemma* (

Game 8.1

) and suppose that Jill expects Jack to defect (d). Her own payoffs from responding with d depend, claims Rabin, on the reasons for which she thinks Jack is about to play d . In effect, her utility payoff from Nash outcome dd will be different depending on whether she thinks that Jack plans to play d because:

(A) *he anticipates a cooperative move (c) from her* and, thus, is rubbing his hands gleefully at the prospect of cheating on her, or

(B) *he expects her to play d also*.

In the latter case, Jill has no reason to be annoyed with Jack. Each expects the other to adopt their dominant strategy (d) and they respond accordingly. However, in case (A) Jill thinks that Jack is intentionally shunning her efforts to achieve a mutually advantageous (or collectively rational) outcome. Jill has reason to be annoyed with him and to think: 'He knows I am taking a risk to our mutual benefit and yet he is not reciprocating, the cad.' This unpleasant thought means that the *same* outcome (dd) may have different utility repercussions for Jill depending on whether her perception of Jack's motivation is given by (A) or by (B) above.

Putting the same case in terms of the material versus the psychological payoffs on [equation \(8.1\)](#)

, in case (A) Jill thinks that Jack's choice of d is 'morally' neutral. Playing d herself, in response, has thus no psychological effect on her. However,

in case (B) she thinks that Jack is willingly and knowingly refusing to return a favour and, therefore, the same outcome (dd) leaves her with the same material payoff, but also with a bitter aftertaste. In terms of total utility, she is less satisfied than under case (A).

Rabin then codifies Jill's perception of Jack's motivation in terms of a combination of her first- and second-order beliefs. The main assumption is that we lose psychological utility when we either fail to return favours or kindness (by performing similar favours or acts of kindness) or fail to punish (even if doing so is costly to us) those who are being nasty. Rabin's model is, ostensibly, founded on the following combination of definitions and assumptions:

Sacrifice: We say that Jill makes a sacrifice, viz. Jack when she intentionally forfeits part of her material utility in order for Jack to receive utility different to that which Jill thinks Jack is 'entitled' to.

Reciprocity: When Jill predicts that Jack is about make a sacrifice on her behalf (see above definition), Jill experiences an urge to reciprocate. If that urge remains unfulfilled, she suffers some psychological loss (i.e. $\Psi < 0$) – see also the previous chapter for a detailed analysis of reciprocity in the context of my theory of solidarity.

Symmetry: The psychological urge to reciprocate is symmetrical in that it applies equally when Jack is expending utility in order (a) to benefit Jill (i.e. increase Jill's utility beyond his 'entitlement'), and (b) to hurt Jill (i.e. reduce her utility below her 'entitlement'). For example, if Jill predicts that Jack will sacrifice utility on her behalf, a failure on her part to reciprocate sets her $\Psi < 0$ *irrespective of whether his sacrifice is intended to hurt her or help her*.

Kindness/nastiness/neutrality: When Jack sacrifices utility in order to boost (diminish) Jill's utility beyond (below) what she is entitled to, he is being *kind* (*nasty*). If his actions do not affect Jill's entitlement, he is being neither kind nor unkind to Jill. He is just *neutral*.

In summary, the kind face of reciprocity emerges if Jill perceives that Jack is prepared to forfeit utility in order to benefit her. Then she is bound to suffer psychological disutility ($\Psi < 0$) if she does not sacrifice some of her own utility in order to benefit him back. Reciprocity's unkind face appears when Jill thinks that Jack's sacrifice is intended to hurt her. Rabin (1993) insists, in a manner reflecting an eye-for-an-eye-tooth-for-a-tooth sort of reciprocity, that Jill suffers psychological losses when she allows Jack to 'get away' with such nastiness. Indeed $\Psi < 0$ unless she sacrifices some utility in order to hurt him back!

The crucial aspect of the above set of assumptions is that Jill's willingness to sacrifice utility in order to affect Jack's utility is made dependent on her perception of their *entitlements* (his *and* hers). Recall that sacrifice here is defined not merely in terms of lost utility but as forfeited utility with the explicit purpose of helping someone achieve higher (in case of kindness) or lower (the case of nastiness) utility *relative to that which she is entitled*. In other words, sacrifice cannot

be defined without a concept of entitlement. If your actions causes you to lose utility, whether this constitutes a sacrifice (as the latter is defined here) or not depends on your perception of: (A) how this sacrifice affects the other's utility, and (B) whether the other was entitled to more or less utility than your action has made possible for her to attain. Before we discuss the obvious question regarding the *origin* of these perceptions of entitlements, I shall first introduce the notion of *fairness equilibrium* that this perspective makes possible.

To illustrate I shall consider two familiar games: *hawk-dove* and the *prisoner's dilemma*.

Table 8.1

lists their original payoffs and then relates the way in which one of the two players' beliefs translate into psychological payoffs. Suppose that A believes that B intends to play *h* in *hawk-dove*. Her psychological payoffs depend on: (i) her first-order beliefs (that is, what she expects B to play); (ii) her second-order beliefs (that is, what she

thinks that B expects *her* to play), and (iii) her own intentions. Let us denote these as follows:

- (i) **AbB: *s*** – ‘A believes that B will play strategy *s*’
- (ii) **AbBbA: *s*** – ‘A believes that B believes that A will play strategy *s*’
- (iii) **A: *s*** – ‘A intends to play *s*’

To see how the rationale in each cell is derived, we look at two such cells. First, we look at the cell corresponding to *hawk–dove* and to A’s belief that B will play *h* (AbB:*h*) because he expects her to play *h* (AbBbA:*h*) when, indeed, she intends to play *h* (A:*h*). In

Table 8.1

the relevant cell reads: ‘B is being nasty. I am nasty back. Thus, $\Psi = \alpha > 0$.’ Why is B being nasty, according to A? Because, given A’s expectations, he plays *h* knowing that A will play *h* too. Given his prediction that A will play *h*, his best response (i.e. his Nash strategy) is to respond with *d*. But he does not! Instead, he replies with *h*.

Such behaviour by B is tantamount to nastiness: By playing *h* in response to her *h*, B is sacrificing a material payoff equal to 2 utils. Given that there is CKR, A interprets B’s strategy as a means to hurt her (i.e. ‘B’s objective *must* be that I get –2 rather than +2’, A thinks to herself). Were she to respond to this situation by playing *d* (which is A’s Nash response to the prediction that B will play *h*), she would be letting him get away with his nastiness. Though it is true that she stands to lose material payoffs if she also plays *h*, there is a psychological boost from ‘*not* letting him get away with it’ ($\Psi = \alpha$, when A plays *h*) and a psychological disutility ($\Psi = -\beta$) if she does.

Notice that the cell we just discussed (the top left-hand cell of

Table 8.1

) corresponds to a *psychological equilibrium*: A’s plans to play *h* because she thinks that A will play B because A thinks that B will predict correctly that A will play *h*. In

Table 8.1

all cells corresponding to such *psychological equilibria* are shaded. Heavier shading is used in

Table 8.1

to mark cells in which there is also a coincidence between: (a) what A expects B to expect her to play, and (b) what A intends to do. In these cases, we have a fully fledged psychological (and Nash) equilibrium in the sense that, not only are A’s first- and second-order beliefs in (psychological) equilibrium but, additionally, we have equilibrium (as in Nash) between A’s beliefs about B’s beliefs and B’s actual choice.

Table 8.1

Player A’s psychological payoffs depending on her beliefs

<i>Hawk–dove</i>				<i>Prisoner's dilemma (Game 8.1)</i>			
<i>Material payoffs</i>		B		<i>Material payoffs</i>		B	
		<i>h</i>	<i>d</i>			<i>D</i>	<i>c</i>
A	<i>h</i>	−2, −2	2, 0	A	<i>d</i>	1, 1	4, 0
	<i>d</i>	0, 2	1, 1		<i>c</i>	0, 4	3, 3
<i>A's beliefs and intentions</i>		AbB:h	AbB:d			AbB:d	AbB:c
AbBbA:h	A:h	B is being nasty. I am nasty back. Thus, $\Psi = \alpha > 0$	B plays his Nash strategy. So do I. Thus, $\Psi = 0$	AbBbA:d	A:d	B is being neutral, playing Nash. So am I. Thus, $\Psi = 0$	B is being kind but I am not. Thus, $\Psi = -\kappa < 0$
	A:d	B is being nasty. I am acquiescent. I let him get away with it. Thus, $\Psi = -\beta < 0$	B plays Nash. Why am I being kind? $\Psi = -\gamma < 0$		A:c	B is being neutral. Why am I am being kind to him? Thus, $\Psi = -\theta < 0$	B is being kind and so am I. Thus, $\Psi = \lambda > 0$
AbBbA:d	A:h	B plays his Nash strategy. He is neutral (neither nasty nor kind). I, on the other hand, am being nasty. Thus, $\Psi = -\varepsilon < 0$	B is being kind but I am not kind in return. Thus, $\Psi = -\zeta < 0$	AbBbA:c	A:d	B is being neutral, playing Nash. I do the same. Thus, $\Psi = 0$	B is being kind but I am not. Thus, $\Psi = -\kappa < 0$
	A:d	B plays his Nash strategy. So do I, responding to his neutrality with neutrality. Thus, $\Psi = 0$	B is being kind. And so am I. Thus $\Psi = \eta > 0$		A:c	B is being neutral. Why am I am being kind to him? Thus, $\Psi = -\theta < 0$.	B is being kind and so am I. Thus, $\Psi = \lambda > 0$

Player A's psychological payoffs in *hawk–dove* and the *prisoner's dilemma* depending on the combination of her (first- and second-order) beliefs and her intended choices

NB. Shaded cells denote an equilibrium between A's first- and second-order beliefs. Cells with heavier shading denote also an equilibrium between what A believes B predicts of her and what she predicts of him.

Let us now concentrate on another cell which does not correspond to an equilibrium: The second row and second column in the *prisoner's dilemma* (see

[Game 8.1](#)

); i.e. the case where AbBbA:d, AbB:c, A:c. In this case, A plans to

cooperate, expects B to cooperate, but thinks that B expects her to defect. The belief that B is cooperating (though fully instrumentally rational) makes her feel that he is making a sacrifice which benefits A; i.e. A thinks that B is being kind. Were she to defect, her material payoff gains (payoff 4 rather than 3) would come at the price of knowingly failing to return B's kindness; thus, A would forfeit psychological utility. By reciprocating B's cooperation (second row, last column), she gains psychological utility. Whether, under the circumstances, A would indeed cooperate or not depends on the relative magnitude of the psychological utils and the material gains from defecting.

The cell that we just examined is inconsistent with an equilibrium in the sense that A predicts that B is not predicting A's intentions correctly. Of course, this does not mean that we should discard it. Why should we assume an equilibrium (i.e. a coincidence of B's beliefs with A's actions) when all combinations of beliefs and actions in

[Table 8.1](#)

are *rationalisable*? As we discovered in the preceding chapters, nothing short of telepathy (the axiom that agents' beliefs are, inexplicably, consistently aligned with all choices) will secure an equilibrium outcome. However, it is not at all clear how such an alignment will occur.

The problem is indeed amplified here because without this form of 'telepathy', which can be more 'politely' called the assumption of *consistently aligned beliefs* (CAB), the game's very strategic structure is ill-defined.

Table 8.2

illustrates this. Suppose we wanted to re-write the *hawk-dove* and the *prisoner's dilemma* games taking into account both the material and the psychological payoffs in

Table 8.1

. It is immediately evident that this task is not as straightforward as it was to transform the *prisoner's dilemma* into

Game 8.2

– see

Section 8.1.1

. On that occasion, the transformation was simple because both material and psychological payoffs were defined exclusively in terms of outcomes. Here, the psychological payoffs are also defined in terms of second-order beliefs. This means that, unless we know the players' beliefs about the beliefs of their opponents, the game cannot even be written down!

To see this, let us assume for simplicity that B's psychology is identical to A's (that is, they share the same parameters $\alpha, \beta, \gamma, \dots$), add their psychological to their material payoffs, and thus re-create the payoff matrices of both games. What is A's overall utility payoff from outcome *hh* in *hawk-dove*? We simply cannot tell unless we know her second-order beliefs. For if she thinks that B expects her to play *h*, then her psychological payoff from outcome *hh* is $+\alpha$. But if, in contrast, she thinks that B played *h* expecting her to play *d*, outcome *hh* costs her psychological utility equal to $-\beta$.

Table 8.2

shows that the utility payoffs cannot be written down with any degree of certainty.

For instance, in the *prisoner's dilemma* players managing to cooperate may receive utility equal to $3 - \kappa$ or to $3 + \lambda$, depending on their second-order beliefs. Cheating players (i.e. those defecting against a cooperative opponent) will, similarly, get utility equal to $4 - \kappa$ or to $4 + \lambda$, again depending on their second-order beliefs. In the same vein, players in the *hawk-dove* interaction will value outcome *dd* differently depending on their second-order beliefs (i.e. their utility equals either $1 - \zeta$ or $1 + \eta$). So, unless we know what they think that

their opponent expects of them, we cannot know their utility from any of the outcomes.

The significance of this cannot be overstated: Game theory has traditionally assumed that rational beliefs are to be extracted from the given structure of the game. When games turn psychological, however, the game's structure is variable and changes drastically as beliefs shift. The gist of this inter-dependence between strategic structure and beliefs is that the former cannot be used in order mechanistically to derive the latter. Therefore, indeterminacy becomes the order of the day as it is clear that without determinate beliefs there can be no unique prediction of what will happen.

All that game theory can do in this instance is to perform its favourite trick: Assume equilibrium and then find it! This is precisely what we do in the second part of

Table 8.2

: By assuming that A takes it for granted that B will predict her choices accurately (and vice versa) we can pin down one psychological payoff per cell. We add this to the

corresponding material payoff and we end up with a single total utility payoff per cell per player.

Table 8.2

Consistently aligned beliefs as a prerequisite for a well-defined game. How the assumption of consistently aligned beliefs (CAB), a form of our third meta-axiom (see

Chapter 1

), became a pre-requisite for well-defined psychological games: (a) The *hawk-dove* and *prisoner's dilemma* games with psychological payoffs: it is no longer possible to define their strategic structure a priori; and (b) Under psychological equilibrium or enhanced CAB (i.e. coincidence of first- and second-order beliefs), the games are well-structured and yield Nash equilibria (known as *fairness equilibria*) which may differ from the original game's Nash equilibria

(a)

	h	d		d	c
h	$-2 + \alpha$ or $-2 - \beta, -2 + \alpha$ or $-2 - \beta$	2 or $2 - \gamma, 0$ or $-\varepsilon$	d	1 or $1 - \theta, 1$ or $1 - \theta$	$4 - \kappa$ or $4 + \lambda, 0$ or $-\theta$
d	0 or $-\varepsilon, 2$ or $2 - \gamma$	$1 - \zeta$ or $1 + \eta,$ $1 - \zeta$ or $1 + \eta$	c	0 or $-\theta, 4 - \kappa$ or $4 + \lambda$	$3 - \kappa$ or $3 + \lambda,$ $3 - \kappa$ or $3 + \lambda$
<i>Hawk-dove</i>			<i>Prisoner's dilemma</i>		

(b)

	h	d		d	c
h	$-2 + \alpha, -2 + \alpha$	$2, 0$	d	$1, 1$	$4 - \kappa, -\theta$
d	$0, 2$	$1 + \eta, +\eta$	c	$-\theta, 4 - \kappa$	$3 + \lambda, 3 + \lambda$
<i>Hawk-dove</i>			<i>Prisoner's dilemma</i>		

The result is a transformed game featuring equilibria (called *fairness equilibria*) that may differ quite sharply from the original Nash equilibria. To see the difference, consider first *hawk-dove*.

Table 8.2b

reveals that, provided $\eta > 1$, A's best reply to *d* is *d* (rather than *h*). Outcomes *hd* and *dh* cease to be equilibria in the *hawk-dove* game, giving their place to the dovish *dd*. The interpretation here is that mutually dovish behaviour is perfectly rational when there are considerable psychological rewards from 'rewarding' rational opponents who resist the temptation to profit at your expense. The other side of this *fairness equilibrium* is that, at the same time, *h* may be a best reply to *h* (provided $\alpha > 2$): As long as players enjoy considerable psychological utility from punishing 'nasty' opponents, they may well get locked into an un-fairness equilibrium (i.e. a never-ending sequence of self-confirming first- and second-order beliefs) which is sustained by the psychological utility from punishing one another's intention to harm.

Turning to the *prisoner's dilemma*, a similar result is arrived at. Once we assume equilibrium, there are two fairness equilibria: Mutual defection is always one (as 1 always exceeds $-\theta$ and, therefore, *d* is always a best reply to *d*). But there is a second one: mutual cooperation; the holy grail of political scientists who have not lost hope that rational cooperation can be rationalised in the context of a one-shot prisoner's dilemma (recall, for instance, Gauthier's efforts from

Chapter 5

). All that is necessary from cooperation to be consistent with equilibrium is that $3 + \lambda > 4 - \kappa$, or that $\lambda + \kappa > 1$. When this is so, the *prisoner's dilemma* becomes a form of

stag-hunt game featuring two distinct fairness equilibria: *dd* and *cc*.

In summary, a psychological feedback from beliefs to preferences adds realism to the analysis but only at the cost of even more indeterminacy!

6

Fairness equilibria: Consider games in which the players' utilities have been augmented with psychological utility (Ψ). Suppose further that these psychological utility functions (Ψ) satisfy the *sacrifice*, *reciprocity* and *symmetry* conditions, as well as the definition of *kindness* above. The Nash equilibria of the resulting game are known as *fairness equilibria*. They are associated with the notion of fairness because the psychological utility underpinning them rises in response to the belief that one is acting in a manner that 'aids' ('harms') an opponent who is making a sacrifice in order to 'aid' ('hurt') one. Otherwise, the Nash idea of 'solving' a game, by focusing on some equilibrium between *beliefs* (first- and second-order in this case) and *actions*, remains intact.

8.2.3 Rabin's formulation of entitlements and kindness functions

This subsection illustrates Rabin's *fairness equilibrium* in the two games already studied above: the *hawk-dove* and the *prisoner's dilemma*. Similarly to our equation (8.1)

, Rabin (1993) supposes that overall utility is the weighted sum of the material payoffs $M(O)$ from outcome O and the psychological payoffs $\Psi(O)$:

$$U_i(O) = (1 - v)M(O) + v\Psi(O) \quad (8.2)$$

Clearly, the higher the value of v the more the player cares about her psychological rewards from a certain outcome, relative to her material payoffs. Utility function (

8.2

) can be re-written for simplicity as (

8.3

) with $\mu = v/(1 - v)$ capturing the relative weight of psychological to material payoffs:

7

$$U_i(O) = M(O) + \mu\Psi(O) \quad (8.3)$$

Kindness/nastiness functions reflecting A's beliefs

A's kindness function toward B, f_A : Rabin defines f_A as a function which varies between -1 and $+1$. If it takes a value between 0 and $+1$, this means that A *believes* that she is being kind to B , given her estimates of what B will do and what she thinks he expects her to do. Similarly, if $f_A < 0$, A thinks that she is being nasty to B . Finally, if $f_A = 0$, A deems that she is being 'neutral'.

B's kindness function toward A, f_B : This is a similar function, also varying between -1 and $+1$, depending again on A's beliefs. In this case, $f_B > 0$ means that A thinks that B is being kind to her (given her estimates of what she expects him to do *and* of what she expects him to predict that she will do). Naturally, if $f_B < 0$, her (first- and second-order) expectations lead her to the prediction that he is being nasty to her. Equality $f_B = 0$ brings to A a feeling that B is being 'morally' neutral to her.

Rabin now defines A 's psychological payoffs in terms of these kindness functions:

$$\Psi_A(O) = f_B(O)[1 + f_A(O)] \quad (8.4)$$

This function embodies the features of reciprocation noted above. Thus, when A anticipates that B is going to be 'kind' (i.e. $f_B(O) > 0$), then her psychological payoffs are positive when she reciprocates (i.e. with $f_A(O) > 0$) while being nasty ($f_A < 0$) turns them negative. Alternatively, if A expects B to be nasty [$f_B(O) < 0$], then only way of making the psychological payoffs non-negative is to make f_A negative (i.e. to be nasty back). Finally, if she anticipates neutrality from B [$f_B(O) = 0$], it makes no psychological difference to her whether she is kind, nasty or neutral to B . However, as both kindness and nastiness require (by definition) a sacrifice of material payoffs from A , she has no reason to make it (since her psychological rewards from it are zero).

The next step in the mathematical representation is to define functions f_A and f_B . They are given below in (

8.5

) and (

8.6

) for each possible outcome. Since each outcome (of a two person game) corresponds to a combination of strategies, s_A for A and s_B for B , the kindness function is defined over these possible strategy pairs:

$$f_A(s_A, s_B) = \frac{\pi_B(s_A, s_B) - e^B(s_B)}{\pi_B^h(s_B) - \pi_B^l(s_B)} \quad (8.5)$$

$$f_B(s_A, s_B) = \frac{\pi_A(s_A, s_B) - e^A(s_A)}{\pi_A^h(s_A) - \pi_A^l(s_A)} \quad (8.6)$$

Kindness is given here as a ratio. The numerator in f_A is simply the difference between A 's estimate of B 's material payoff and his 'entitlement' given that he plans to play s_B [that is, $e^B(s_B)$]. So this measures A 's kindness to B . The role of the denominators is to keep the values of f_A and f_B within the required bound of $(-1, +1)$ and they are the difference between player i 's maximum and minimum material payoffs when he or she plays s_i : $\pi_i^h(s_i) - \pi_i^l(s_i)$

Let us now look at the *hawk-dove* and suppose that, for some reason, A expects B to play d . A knows that, depending on whether she chooses d or h , B 's largest possible material payoff is $\pi_B^h(s_B = d) = 1$ and his minimum is $\pi_B^l(s_B = d) = 0$. Clearly, the denominator of (

8.5

) equals 1. Suppose further that A plans to play h . In that case, $f_A(s_A = h, s_B = d) = [0 - e^B(s_B = d)]/1$. Evidently, in this situation, A is either nasty ($f_A < 0$) or, at best, morally neutral ($f_A = 0$) toward B , depending on B 's 'entitlement'. Note that, in her own mind, she is being nasty if she thinks that B was entitled to something more than payoff 0 given that he played d (i.e. if $e^B(s_B = d) > 0$). But if she thinks that a d -playing B is entitled to nothing, she believes that her h -choice was morally neutral (in the sense that it did not deny B of any entitlement).

Similarly, suppose that she thinks that B expects h of her and plans to play d in response. Is he being kind, nasty or neutral towards A ? Let's find out the answer as reported by (

8.6

). If A is expected to play h , B knows that her highest and lowest payoffs are 2 and -2 respectively. Thus, the denominator equals $\pi_B^h(s_B = h) - \pi_B^l(s_B = h) = 2 - (-2) = 4$. As for the numerator, it equals $\pi_A(s_A = h, s_B = d) - e^A(s_A = h) = 2 - e^A(s_A = h)$. Suppose that A believes that, when she plans to play h in the expectation that B will play d , she is entitled to get material payoff 2 (i.e. $e^A = 2$). In that case, she thinks that B is being neutral since $f_B = 0$.

Finally, to complete the model, the mathematical representation of Rabin's idea regarding entitlements is implied by its definition (see above). To illustrate their calculation, suppose that A is involved in a two-person game in which if she plays her first of three strategies there are three possible outcomes, depending on which of the three strategies available to him her opponent chooses. The following payoffs come from Game 6.4 in the previous chapter: $(-2, -2)$, $(2, 0)$, $(4, -1)$. What is A entitled to, according to Rabin, when she chooses this particular strategy?

According to the above definition, $e_A(s_A) = 3$. To see this, first we note that of the three outcomes associated with A 's choice the first one is dominated and thus discarded. By this we mean that $(-2, -2)$ is clearly worse for *both* players than either $(2, 0)$ or $(4, -1)$ – and in this sense it is dominated – and therefore does not count in the computation of A 's entitlement. Of the remaining two outcomes, neither dominates the other since there is no way one of them could be discarded without one player objecting. Of those two remaining outcomes, A 's possible material payoffs equal either two or four utils. The average of these is three and this is, according to Rabin what A is entitled to if she plays this strategy. Of course, this being an average, it is obvious that A will either get

more than she is entitled to (i.e. four) or less (i.e. two). But such is life. We seldom get what we deserve in life. We are either too far ahead or struggling to catch up!

With the entitlements of both players $e^A(\cdot)$ $e^B(\cdot)$ fully computed for each of their strategies,

equations (8.5)

and (

8.6

) can now be computed for each outcome. These values are then put back into

equation (8.4)

to find the psychological payoffs per outcome before imputing these into (

8.3

). At that stage the game has been totally transformed and a new payoff matrix is derived depicting the players' overall payoffs (material and psychological). The Nash equilibria of this transformed game are the *game's fairness equilibria* – as defined previously.

Table 8.3

begins with the 'standard' material payoff representations of the two static games.

Then it notes for each of player i 's strategies the following: (a) π_i^h – i.e. i 's maximum payoff possible for the given strategy, (b) π_i^ℓ – i.e. i 's minimum payoff possible, and (c) e^i – i.e. i 's entitlement if indeed he/she chooses that strategy.

For instance, in the *hawk-dove*, the most B can expect when choosing h is 2 utils while the least is -2 . As for his 'entitlement', we recall that Rabin demands of us that, *before* we average out B 's possible payoffs, we discard any 'dominated' outcome corresponding to B playing h . There are two possible outcomes when B plays h : $(-2, -2)$ and $(2, 0)$. Clearly $(-2, -2)$ is dominated by $(0, 2)$ in the sense that neither player would object if some adjudicator forced them to trade the former for the latter. Thus, we discard $(-2, -2)$. This leaves B with only one possible payoff (for the purposes of computing his entitlement) when he plays h : 2 utils. Thus, according to Rabin, if B chooses strategy h , he is 'entitled' to 2 utils – i.e. $e^B(h) = 2$.

Let us perform the same computation once more, only this time for when A plays d in the *prisoner's dilemma*. Since she is 'defecting', A 's highest possible payoff is 4 and her lowest 1; i.e. $\pi_A^h(d) = 4$ and $\pi_A^\ell(d) = 1$. Her entitlement? Just as before, we investigate whether there is a dominated outcome corresponding to A playing d . The two potential outcomes are: $(4, 0)$ and $(1, 1)$. Neither dominates the other, in the sense that, if the players were to be forced away from one and onto the other, one of them would protest. Thus, Rabin insists, A 's entitlement $e^A(d)$ equals the average of her two potential payoffs 4 and 1. That is, $e^A(d) = 2.5$.

In this manner we compute all values of π^h , π^ℓ and e for both players and all their strategies. Next we input these values into expressions (

8.5

) and (

8.6

) to compute the levels of kindness/nastiness that one shows to the other for each combination of strategies (and thus for each outcome). Suppose, for example, that A plays h and B responds with d in *hawk-dove*. How kind/nasty is A being to B ? In other words, is f_A positive or negative? From expression (

8.5

) we compute it as a ratio between two differences. The numerator is the difference between B 's payoff (when A plays h and he plays d) and his entitlement given that he chose d (the numerator thus equals $0-1/2$). The denominator equals the difference between B 's maximum and minimum potential payoffs when choosing d

(i.e. $1 - 0$). In conclusion, $f_A = -1/2$ and A is deemed to be mean (or nasty) toward B .

Once f_A and f_B have been computed for all outcomes, it is straightforward to utilise expression (

8.4

) in order to compute both players' psychological payoffs for each outcome. To continue with the example of the above paragraph, suppose again that in *hawk-dove* A plays h and B d . What are A 's psychological payoffs? According to expression (

8.4

), for each outcome we add 1 to the corresponding value of f_A and multiply what we find with the corresponding value of f_B . This product gives us A 's psychological payoff: When A and B play h and d respectively, $f_A = -1/2$, $1 + f_A = 1/2$, and $f_B = 0$. Thus, A 's psychological payoff equals zero [$\Psi_A = f_B (1 + f_A) = 0$].

What is the meaning of this finding? Both A and B played their Nash best replies (recall that, in *hawk-dove*, h is the best reply to d and vice versa). However, A was branded 'nasty' by Rabin's formulation ($f_A < 0$). And yet she lost *no* psychological utility from this imputed 'nastiness' [$\Psi_A = 0$]. Why? The reason is that, in playing d , B was *not* making *any* sacrifice on A 's behalf. Thus, A had no moral obligation to be kind to him (that is, to sacrifice some of her material payoffs on his behalf). Remember, for Rabin nastiness leaves a bitter aftertaste in the nasty player's mouth only if her opponent was being kind. Here, B was being neither kind nor unkind. He was simply playing his Nash best reply and A felt no obligation to make a sacrifice on his behalf. In fact she felt no psychological effects whatsoever as a result of her mild 'nastiness' toward him.

With the computation of psychological payoffs completed, we can now add them to the original (or material) payoffs, according to expression (

8.2

). The result is the psychological transformation which yields the final payoff structures at the bottom of

Table 8.3

Once the psychological effects of fairness have been incorporated, the game's strategic structure is transformed drastically. Of course, the extent of this change depends on how much importance players attach to their psychological 'side' relative to hard-nosed considerations of material payoffs. One might, for example, feel bad when double-crossing a friend or foe but, at the same time, place little emphasis on this ill feeling relatively to the appreciation of the material benefits from such treachery. In this model, it is parameter ν (or μ) that captures the value of psychological payoffs (relative to the material ones). [Recall that as ν tends to 1, or equivalently μ tends to infinity, psychological payoffs tend totally to over-shadow material ones. And vice versa as ν tends to zero.]

Starting with *hawk-dove*, nothing changes as long as the relative 'weight' of the

players' 'psychology' falls below a certain threshold ($v < 0.57$). Once it exceeds that threshold, the game is transformed utterly and the two original pure strategy Nash equilibria (A plays h , B plays d or B plays h and A plays d) cease to exist, giving their place to a unique (fairness) equilibrium in which both players opt for d . The point here is that, as long as psychological payoffs matter sufficiently, the only possible equilibrium is one in which one believes that the other is making a sacrifice on one's behalf (in playing d , as opposed to reaping maximum material payoffs) and therefore feels that, if this sacrifice is not reciprocated, the

loss of psychological utility would be greater than the gain of material utility from playing h .

The startling aspect of the transformed *hawk-dove* is reinforced when players place even more importance to the psychological aspects of the game. For if $v > 0.8$, a *second* fairness equilibrium comes to light: in addition to dd , hh is an equilibrium too in the context of which each plays aggressively in the full knowledge that the other will do likewise, with the end result that both will forfeit 2 material utils. This is an equilibrium oozing mutual nastiness; a kind of eye-for-an-eye situation sustained by the urge each feels to inflict 'pain' on a player who is trying to hurt her because he predicts (correctly) that she will inflict pain on him because she thinks (correctly) that he will inflict pain on her... *ad infinitum*.

Things are also different in the case of the *prisoner's dilemma*. The original mutual defection equilibrium (dd) survives independently of the relative weight of psychology. The idea is that, the emphasis on psychological losses makes no difference here since the decision of an opponent to defect (unlike the decision to play h in *hawk-dove*) involves no sacrifice and thus does not give a player an urge to reciprocate. So, mutual defection is a fairness equilibrium in the sense that one player is morally neutral to the other. What *does* change however is that, provided psychology matters sufficiently (i.e. $v > 0.66$), mutual cooperation (cc) becomes an additional equilibrium. This is very similar to the dd fairness equilibrium in *hawk-dove*, as it is based on the mutual expectation that the other is making a sacrifice on your behalf because she expects you to make a similar sacrifice too, in the belief that you are anticipating her sacrifice ... *ad infinitum*. Interestingly, as long as $v > 0.66$, the *prisoner's dilemma* is transformed into a kind of *Stag-Hunt* game with two (fairness) equilibria: cc and dd .

8.2.4 An assessment of Rabin's entitlements

Three things stand out in Rabin's 1993 theory. The first is that reciprocation matters, but the precise way in which psychological payoffs are generated through *reciprocation* is potentially controversial. Reciprocity is a common feature of most normative theories of action in the sense that people seem more likely to be influenced by a norm when they interact with others who are similarly influenced. So if a norm dictates 'kindness' then it is indeed more likely that people will follow this dictate when they expect others to.

This is what much of the experimental evidence points to and so this is an important feature of the theory, which we return to below. Rabin however also makes it more likely that someone will behave 'nastily' when they expect 'nastiness' from others and this seems rather less plausible as a general proposition. An 'eye-for-an-eye' is, after all, but one piece of folk wisdom. 'Turning the other cheek' is another that would go directly against this kind of reciprocation of 'nastiness'. Old Testament versus the New, so to speak; and it is not obvious that either could stand for the general case.

Secondly, the nature of what is being reciprocated seems rather special. 'Kindness' is plainly one aspect of behaviour that people often value, but it is not the only one. The claims of 'justice', 'goodness' and 'honour', which can all come in a variety of forms, seem just as strong.

Table 8.3

The derivation of Rabin's fairness equilibria using only material payoffs

Hawk–dove				Prisoner's dilemma			
Original pay-offs	B			Original payoffs	B		
		<i>h</i>	<i>d</i>			<i>d</i>	<i>c</i>
	<i>h</i>	−2, −2	2, 0		<i>d</i>	1, 1	4, 0
A	<i>d</i>	0, 2	1, 1	A	<i>c</i>	0, 4	3, 3

Rabin's derivation of players' maximum/minimum payoffs and of their entitlements *per strategy*

	B: <i>h</i>	B: <i>d</i>			B: <i>d</i>	B: <i>c</i>			A: <i>d</i>	A: <i>c</i>
π^h_B	2	1		π^h_A	4	3		π^h_B	4	3
π^d_B	−2	0		π^d_A	1	0		π^d_B	1	0
e^B	2	½		e^A	2.5	1.5		e^A	2.5	1.5

Computing the kindness/nastiness functions f_A and f_B – see equations (8.5), (8.6)

f_A	B: <i>h</i>	B: <i>d</i>		f_B	B: <i>h</i>	B: <i>d</i>		f_A	B: <i>d</i>	B: <i>c</i>		f_B	B: <i>d</i>	B: <i>c</i>
A: <i>h</i>	−1	−½		A: <i>h</i>	−1	0		A: <i>d</i>	−½	−½		A: <i>d</i>	−½	½
A: <i>d</i>	0	½		A: <i>d</i>	−½	½		A: <i>c</i>	½	½		A: <i>c</i>	−½	½

Computing A's/B's psychological payoffs: $\Psi_A = f_B(1+f_A)$ and $\Psi_B = f_A(1+f_B)$ – see equation (8.4)

$1+f_A$	B: <i>h</i>	B: <i>d</i>
A: <i>h</i>	0	½
A: <i>d</i>	1	1½

$\Psi_A = f_B(1+f_A)$	B: <i>h</i>	B: <i>d</i>
A: <i>h</i>	0	0
A: <i>d</i>	−½	¾

$1+f_B$	B: <i>h</i>	B: <i>d</i>
A: <i>h</i>	0	1
A: <i>d</i>	½	1½

$\Psi_B = f_A(1+f_B)$	B: <i>h</i>	B: <i>d</i>
A: <i>h</i>	0	−½
A: <i>d</i>	0	¾

$1+f_A$	B: <i>d</i>	B: <i>c</i>
A: <i>d</i>	½	½
A: <i>c</i>	1½	1½

$\Psi_A = f_B(1+f_A)$	B: <i>d</i>	B: <i>c</i>
A: <i>d</i>	−¼	¼
A: <i>c</i>	−½	¾

$1+f_B$	B: <i>d</i>	B: <i>c</i>
A: <i>d</i>	½	1½
A: <i>c</i>	½	1½

$\Psi_B = f_A(1+f_B)$	B: <i>d</i>	B: <i>c</i>
A: <i>d</i>	−¼	−½
A: <i>c</i>	¼	¾

The 'psychologically' transformed games in equilibrium

	<i>h</i>	<i>d</i>
<i>h</i>	−2, −2	2, −½μ
<i>d</i>	−½μ, 2	1 + ¾μ, 1 + ¾μ

Hawk–dove

	<i>d</i>	<i>c</i>
<i>d</i>	1 − ¼μ, 1 − ¼μ	4 + ¼μ, −½μ
<i>c</i>	−½μ, 4 + ¼μ	3 + ¾μ, 3 + ¾μ

Prisoner's dilemma

Three potential fairness equilibria depending on the relative value of psychological payoffs [v; recall that $\mu = v/(1-v)$]

Two potential fairness equilibria depending on the relative value of psychological payoffs [v; recall that $\mu = v/(1-v)$]

Finally, granted that kindness is what is being reciprocated in a particular social setting, it is not obvious that it will always be identified in this precise way. The most obvious cause for concern here is Rabin's identification of how people

perceive their 'entitlements'. Again, it seems more plausible, at least on the basis of the anthropological record, that people's ideas regarding 'entitlements' depend on theories of justice that in turn vary across time and space.

It is not hard, however, to see how each of these points could be met while retaining the same basic model. For instance, the relation between the f functions in the psychological component of people's utility functions could be changed (i.e. a different mathematical form for

[equation \(8.4\)](#)

). Likewise, the f functions could be defined in terms of the extent to which each person has chosen the act which maximises whatever social welfare function best represents the shared view of what is just; and so on. We supply an illustration of this sort in the following subsection. All changes of this sort nevertheless beg a question of where these ideas regarding what is worthy in an action come from. The suggestion that different assumptions could be made merely highlights this point. We need, in short, a theory of norm formation. One natural place to go for this is ... history; and I pursue this thought briefly in

[Section 8.3](#)

For now, I conclude this discussion with a comment of how Rabin and other versions of psychological games have changed game theory. The highlight of the preceding analysis is the thought that, once psychological utilities enter the scene, the theorist

needs to know the character of people's beliefs about each other's actions before payoffs are calculated and alternative strategies assessed. To get to a unique utility assessment of an outcome, we must assume an equilibrium of beliefs. This turns what used to be a simple unidirectional system of causation in game theory, running from utilities to rational beliefs to equilibrium, into a form of circularity. This is especially disturbing when the requirement that the beliefs be in equilibrium do not typically produce a unique set of beliefs, as seems to be the case in psychological games. To use the apparatus of game theory to predict what rational people will do, we need to know what beliefs *actually* obtain. But if one knows this, then the apparatus of instrumental rationality is no longer really needed to explain how people act.

It is true, of course, that action can be instrumentally rationally reconstructed once the beliefs are known, but knowing the equilibrium beliefs about action is enough to predict what actions will be taken and it seems almost simpler to say that people's actions have been guided by the prevailing norm. This is the key aspect of psychological games: payoffs and outcomes are both norm-driven.

There is another way of appreciating what has changed in this chapter. Since indeterminacy has plagued game theory from the start of this book, it may seem that the indeterminacy of psychological games adds nothing to the argument. However, until this chapter the indeterminacy has suggested a weakness in the scope of the instrumental model of rationality, rather than a fundamental flaw. The model itself still had value once some theory of belief formation was grafted on (e.g. through a combination assuming a bounded form of this rationality and beliefs that are generated through an evolutionary process). So, one would have to concede that something else was needed to explain action, but it still made sense to talk about people acting so as to satisfy their preferences. In this chapter, the contrast has been most marked because the indeterminacy goes to the heart of

the model. 'Preferences' are not given independently of beliefs and the indeterminacy of belief yields indeterminate preferences, so talk of acting on preferences becomes difficult to sustain.

8.2.5 An alternative formulation linking entitlements to intentions

Suppose entitlements depend on intentions. In other words, if Jill intends good (bad) things to happen as a result of her actions, then Jack believes that she deserves more (less). Consider any static game between A and B . Suppose that, having predicted that A will choose strategy s_A , B chooses to respond with strategy s_B . Let $E_B(s_B)$ denote our alternative definition of B 's entitlement [juxtaposed against Rabin $e^B(s_B)$]. To ensure that $E_B(s_B)$ depends on the combination of B 's choice (s_B) *and* his intentions (as opposed to just the former), the following must hold if $E_B(s_B)$ is to be non-zero:

(a) B must be sacrificing utility: i.e. $\pi_B(s_A, s_B) < \pi_B^n(s_A)$; where, $\pi_B(s_A, s_B)$ is B 's payoff from choosing s_B (when he expects A to play s_A) and $\pi_B^n(s_A)$ is B 's payoff from choosing his *best reply* strategy in response to s_A , **and**
 (b) A must benefit, or lose out, from B 's sacrifice: i.e. $\pi_A(s_A, s_B) > \pi_A^n(s_A)$ in the case where B is being kind A , or $\pi_A(s_A, s_B) < \pi_A^n(s_A)$ when he is nasty; where, $\pi_A(s_A, s_B)$ is A 's payoff when the two players choose strategies s_A and s_B and $\pi_A^n(s_A)$ is A 's payoff from choosing s_A when B plays his best reply strategy to s_A .

As long as (a) and (b) hold, A must think that B is entitled to a payoff in excess of (less than) $\pi_A^n(s_A)$ by virtue of his kindness (nastiness) to her. In other words, A must feel that B deserves to get something more (less) than what he could have expected under normal Nash-like, best-reply, play. How much more (less)? An obvious (and plausibly 'fair') answer would be that B deserves to benefit (hurt) to a degree proportional to (i) the benefit (loss) he has bestowed upon A , and (ii) to the magnitude of his own sacrifice. Finally, note that if (a) does not hold, then B deserves neither more

nor less than what he will get from normal Nash-like play.

The following is one possible specification for $E_B(s_B)$ satisfying the above requirements:

$$E_B(s_B) = \frac{[\pi_A(s_A, s_B) - \pi_A^n(s_A)]R^B(s_A) + \pi_B^n(s_A)R^A(s_A)}{R^A(s_A)} \times \left| \frac{\pi_B(s_A, s_B) - \pi_B^n(s_A)}{R(s_A) + R(s_A)} \right| \quad (8.7)$$

where $R(s_A)$ is the range of A 's payoffs when she plays s_A ($\max\{\pi_A(s_A)\} - \min\{\pi_A(s_A)\}$) and $R(s_A)$ is B 's range of payoffs when A plays s_A ($\max\{\pi_B(s_A)\} - \min\{\pi_B(s_A)\}$).

Note that, from A 's perspective, (9.6) makes B 's entitlement proportional to the absolute magnitude of *his* sacrifice (relative to the range of both players' payoffs when A plays s_A), to *her* resulting benefit or loss (relative to *her* range of payoffs when she plays s_A) and, finally, to *his* payoffs were he selfishly to stick to his best reply strategy. When B is making no sacrifice one way or another, A 's normative commitment to *his* welfare vanishes; i.e. she does not think that B is *entitled* to anything.

Table 8.4

A 's estimate of B 's entitlements according to equation (8.7)

Hawk–dove				The prisoner's dilemma			
		AbB:h	AbB:d			AbB:d	AbB:c
E_B	AbBbA:h	−2/3 (2)	0 (1/2)	E_B	AbBbA:d	0 (2 1/2)	2/3 (1 1/2)
	AbBbA:d	0 (2)	1 (1/2)		AbBbA:c	0 (2 1/2)	4/3 (1 1/2)

Table 8.4

gives A 's perception of B 's entitlements corresponding to: B 's choice of strategy ($AbB:s_B$), and A 's perception of B 's intention. Note that the latter perception derives from A 's second-order belief ($AbBbA:s_A$); e.g. when $AbBbA:h$ and $AbB:h$, A thinks that B is making a sacrifice in order to hurt her. For why else would he be playing h when he expects her to play h too? Surely, his Nash best reply to her h is d which must mean, A concludes, that in playing h he is deviating from his Nash best reply in order to make her suffer. So, when $AbBbA:h$ and $AbB:h$, A estimates that he is entitled to utility of −2/3.

By contrast, if A thought that the reason why B is about to play h ($AbB:h$) is his belief that she will play d ($AbBbA:h$), then A no longer thinks that B is trying to hurt her. She simply interprets his (predicted) intention to play h as a Nash best reply (and, thus, morally neutral) action. In this case, therefore,

Table 8.4

reports that A believes that B is entitled neither to positive nor to negative material payoffs: he is morally neutral and therefore deserves neither to be helped nor to be harmed by her. Notice that Rabin's formulation makes no such distinction (Rabin's entitlements are in brackets): According to Rabin, A thinks that B is entitled to payoff 2 (his material payoff from the pure strategy Nash equilibrium favouring him) regardless of her interpretation of his intentions. We believe that expression (

9.6

) is much better tuned into the rationale of fairness equilibria.

To see this better, suppose that A expects B to play d as a best reply to her own h (i.e. because $AbBbA:h$). Again A thinks that B is *not* entitled to her benevolence. But, if she thinks that he is playing d in order to help *her* (i.e. when $AbBbB:d$) she thinks that B is entitled to payoff 1; i.e. to a gain greater than payoff 0 which is

proportional both to *her* gain and to *his* sacrifice.

Turning to the *prisoner's dilemma*, first we note that Rabin's specification of *B*'s entitlement is counter-intuitive: *B* is entitled, on the grounds of fairness, to a greater payoff when he defects than when he cooperates (i.e. payoff 1 when he defects and 0 when he cooperates). This is simply unsustainable. In contrast, our specification is such that a defecting *B* does not deserve anything (either

positive or negative), since he is not making any sacrifices either to benefit or to hurt *A*.

8

Indeed, whenever *B* is choosing a dominant (or, more generally, a Nash best reply) strategy he is being, by definition, kindness-neutral and, consequently, *A* 'owes' him nothing (either positive or negative). Entitlements come into play in the *prisoner's dilemma* only when *B* deviates from his dominant strategy and cooperates. In that case, his entitlement is always positive (since his cooperation always benefits *A*) and greatest when *B* is expecting *A* to cooperate too.

9

We now need to define alternative functions to Rabin's (

8.5

) and (

8.6

) so as to measure the kindness/nastiness shown by one player to another given their first- and second-order expectations. We specify *A*'s kindness function to *B* (f_A) so that it takes a positive value when *A* is kind to *B*, a negative value when she is being nasty to him and a zero value when she is being neither kind nor nasty to him.

Re-defining *A*'s kindness/nastiness – expression (8.8)

$$f_A(s_A, s_B) = \begin{cases} 0 & \text{when there exists an alternative strategy } s_A^* \text{ such that} \\ & \pi_B(s_A^*, s_B) > \pi_B(s_A, s_B) > \pi_B^n(s_A) \text{ and} \\ & \pi_A(s_A, s_B) \leq \pi_A(s_A^*, s_B) < \pi_A^n(s_A) \\ \text{OR} \\ & \pi_B(s_A^*, s_B) < \pi_B(s_A, s_B) < \pi_B^n(s_A) \text{ and} \\ & \pi_A(s_A, s_B) \leq \pi_A(s_A^*, s_B) < \pi_A^n(s_A) \\ \left| \frac{\pi_B(s_A, s_B) + R_B(s_B)}{E_B + R_B(s_B)} \right| & \text{when } \pi_B(s_A, s_B) > \pi_B^n(s_A) \\ \text{and } \pi_A(s_A, s_B) < \pi_A^n(s_A) \\ - \left| \frac{\pi_B(s_A, s_B) + R_B(s_B)}{E_B + R_B(s_B)} \right| & \text{when } \pi_B(s_A, s_B) < \pi_B^n(s_A) \\ \text{and } \pi_A(s_A, s_B) < \pi_A^n(s_A) \end{cases}$$

(where R_B is the range of *B*'s payoffs when he chooses strategy s_B)

Expression (8.8) replaces (

8.5

) and offers a more complicated 'theory' of what constitutes kindness/nastiness. The first line of expression (8.8) demands that players think of acts as kind/nasty only if they are efficient. If *A* intends to be nice to *B* by choosing non-Nash strategy s_A but, meanwhile, there exists another strategy s_A^* which would have benefited *B* at no extra cost to her, then *A* is deemed irrational rather than kind. Thus *A*'s kindness function becomes zero. Similarly, when *A* wants to hurt *B*. If her choice of strategy is 'inefficient', her nastiness function is, again, set equal to zero.

10

The second line specifies that *A*'s kindness to *B*, when *B* is sacrificing utility to help her, is a positive function of the proportion of *B*'s entitlement that *A*'s choice allows him

to enjoy. Finally, the last line suggests that, when B is hurting A at a cost to himself, A's nastiness to him is a function of the extent to which A's choice inflicts on B the loss that he deserves (or that she is 'entitled' to inflict upon him).

Table 8.5

A's kindness/nastiness to B according to expression (8.8)

<i>Hawk–dove</i>			<i>The prisoner's dilemma</i>		
f_A	$AbB:h$	$AbB:d$	f_A	$AbB:d$	$AbB:c$
AbBbA:h	$-3/5 (-1)$	$0 (-1/2)$	AbBbA:d	$0 (-1/2)$	$-2/3 (-1/2)$
AbBbA:d	$0 (0)$	$1 (1/24)$	AbBbA:c	$2 (1/2)$	$9/10 (1/2)$

(Rabin's values in brackets).

Table 8.5

presents the values of f_A in our two games depending on A's first- and second-order beliefs, as given by expression (8.8):

Note that (unlike Rabin, 1993) no unkindness is involved when A thinks that B is playing some pure Nash strategy. The reason is that playing Nash involves no sacrifice on B's part and, therefore, it cannot possibly incite (on the strength of reciprocity) any sacrifice from A. Put differently, from a psychological point of view, mutual Nash play is tantamount to kindness-neutrality. Indeed A's kindness (nastiness) surfaces, i.e. $f_A > 0$ (or $f_A < 0$), only when A is acting in a manner that furnishes B with a payoff greater to (less than) he would have expected under Nash play.

Turning to our two games again, we note that in *hawk–dove* player A can show nastiness only to a B whom she expects is playing *h* in order to hurt her.

¹¹

Moreover, in the *Prisoner's dilemma*, no nastiness is involved when players choose to defect (again in sharp contrast to Rabin).

¹²

We are now ready to re-define the psychological payoffs by replacing Rabin's expression (

8.4

) with expression (

8.9

) below. This is a simple yet effective way of capturing A's psychological payoffs (Ψ_A) as the product of f_A and f_B – where the latter is computed by an expression very similar to (8.8). A's overall utility is still given by expression (

8.4

), only this time the psychological component of the player's utility (Ψ_A) is given by our alternative formulation above:

$$U_A(s_A, s_B) = \pi_A(s_A, s_B) + \mu \Psi_A(s_A, s_B) \text{ where}$$

$$\Psi_A(s_A, s_B) = f_A(s_A, s_B) \times f_B(s_A, s_B) \quad (8.9)$$

and μ is, as before, the weight placed by A on her psychological utility (relative to her material utility).

As before, (

8.9

) confirms that, when A anticipates kindness (nastiness) from B, she loses psychological utils if she fails to reciprocate that kindness (nastiness). Note however that, in a manner reflecting Rabin's discussion better than his own formulation, the utility function above takes different values depending on A's second-order beliefs.

¹³

Let us now re-write in

Table 8.6

the overall payoffs *in equilibrium* (recalling once more the crucial point that, out of equilibrium, psychological games are ill-defined) for games *hawk–dove* and *the prisoner's dilemma*.

Table 8.6

The transformed games under the re-defined psychological payoffs

<i>Hawk–dove</i>				<i>The prisoner's dilemma</i>			
		<i>B</i>				<i>B</i>	
		<i>h</i>	<i>d</i>			<i>d</i>	<i>c</i>
<i>A</i>	<i>h</i>	$-2 + 0.36\mu,$ $-2 + 0.36\mu$	$2, 0$	<i>A</i>	<i>d</i>	$1, 1$	$4 - 1.33\mu,$ -1.33μ
	<i>d</i>	$0, 2$	$1 + \mu, 1 + \mu$		<i>c</i>	$-1.33\mu,$ $4 - 1.33\mu$	$3 + 0.81\mu,$ $3 + 0.81\mu$

Three potential fairness equilibria
depending on the relative value of
psychological payoffs
[v ; recall that $\mu = v/(1 - v)$]

Two potential fairness equilibria
depending on the relative value of
psychological payoffs
[v ; recall that $\mu = v/(1 - v)$]

The fairness equilibria that result are similar in structure to those following Rabin's transformation. However, there is one important analytical difference with Rabin's model: *Our transformation, unlike Rabin's, is such that Nash play is psychologically neutral (i.e. has no psychological effects)*. Moreover, they are consistent with the idea that players' perceptions of entitlements, and thus of fairness, depend on their perceptions of the motives behind their opponents' actions.

To see the point about the psychological neutrality of Nash equilibria, consider the original pure strategy Nash equilibria of both games (*hd* & *dh* in *hawk–dove* and *dd* in the *prisoner's dilemma*): Our transformation leaves the associated payoffs intact. The reason is that, in our case, *Nash play is, by definition, psychologically neutral for players* (i.e. their psychological payoffs are zero). If *cc* in the *prisoner's dilemma* and *dd/hh* in *hawk–dove* become fairness equilibria, it is because they reward players with significant psychological rewards (such that the material incentive to play Nash is overcome). By contrast, Rabin's transformation assigns, wrongly we think, non-zero psychological payoffs to Nash equilibria.

Now, this difference is not a mere technicality as it affects our interpretation of fairness equilibria. For example, consider mutual defection in the *prisoner's dilemma*. Both our transformation and Rabin's report that *dd* is a fairness equilibrium (independently of the players' relative valuation of material and psychological payoffs). However, Rabin is forced to insist that *dd* must necessarily be a mutual nastiness equilibrium. In our case this is not so: *Mutual defection is a mutual kindness-neutral equilibrium*.

We think this is important because our re-configuration of fairness equilibrium throws a bridge between the concept of fairness equilibria examined in this section

and the idea that games are regulated by a sense of justice springing out of the expectations that people will (or ought to) comply with the current conventions for playing the game.

8.2.6 Conclusion

In this section, we re-defined players' entitlements so as to reflect not only their actions

but also their intentions. Moreover, we postulated that players are only entitled to 'something' when they deviate from their Nash best replies either to benefit or to harm their opponents. Under these assumptions, we showed that mutual defection in the prisoner's dilemma *cannot* be a mutual-nastiness equilibrium. Rather, it occurs when players are locked in expectations of mutual kindness-neutrality. More generally, in an equilibrium between first- and second-order beliefs, a game's original (pure strategy) Nash equilibria come with zero psychological payoffs (unlike in Rabin, 1993). However, they may not be fairness equilibria (e.g. when $\mu > 0.5$ in *hawk-dove*) because other competing outcomes (e.g. mutual dovishness) may offer players a positive inner glow which the original Nash equilibria cannot match.

8.3 Psychology and evolution

8.3.1 On the origins of normative beliefs: an adaptation to experience

While it is true that psychological game theory *does* explain non-Nash cooperation in games like *hawk-dove* and the *prisoner's dilemma*, it cannot explain why some groups are drawn to cooperation while others are not – see

[Chapters 10](#)

and

11

for some empirical evidence to that effect. One possible explanation is that perceived entitlements adapt to the players' past payoffs. In other words, to return to the argument in the previous section, that 'entitlements' should not be treated exogenously, one way of endogenising them is to appeal to an evolutionary process.

Sugden (1986) argued that predictive beliefs have a tendency to become normatively charged. Echoing David Hume, he suggested that agents find it hard to accept that the convention which determines their behaviour could have been otherwise (even though it might easily have been), so people develop normative reasons to support the convention. It is not only a coordinating device, it embodies ideas of 'justice', 'fairness', etc. Interestingly, if this is how the normative beliefs of both advantaged and disadvantaged groups evolve in games featuring asymmetrical Nash equilibria, like the *hawk-dove* game, the observation of stable discrimination patterns ceases to be a puzzle. Groups that are disadvantaged by some arbitrary characteristic (e.g. being black or women) would develop humbler entitlement expectations as compared with those who are advantaged. As a result, with a sufficiently lower set of perceived entitlements, the conflictual outcome *hh* could cease to be a fairness equilibria and *dd* could become one. By contrast, advantaged players with higher perceived entitlements may not be

spared *hh* and may find themselves locked into a mutual *hawkish* equilibrium with players of the same colour. Why don't they 'evolve out' of these normative expectations if the latter cause them to fight each other at a great cost? The simple evolutionary answer is that such normative beliefs, despite causing much conflict between the 'strong', reward them amply in meetings with the 'weaker' players.

8.3.2 On the origins of normative beliefs: The resentment-aversion versus the subversion-proclivity hypotheses

Sugden (2000) offers a rather different account of how conventions (or mere empirical regularities) come to motivate through affecting player's payoffs. He proposes a psychological equilibrium that he calls a *normative expectations equilibrium* (NME) which is similar to Rabin's *fairness equilibrium*, but which dispenses with any account of the character of the norms that affect behaviour. Instead, it is enough, rather like the early discussion of how second-order beliefs motivate in

[Section 8.2](#)

, that people expect someone to behave in a particular way (whatever it is) for that person to incline towards that action on psychological grounds. This psychological

mechanism seems to be traced to its evolutionary role in conflict avoidance and is captured by the idea that humans are averse to the resentment of others. This is his *resentment hypothesis*.

Sugden's fairness as founded on his resentment-aversion hypothesis

Fairness: In equilibrium, person *A* is *fair* towards person *B* as long as *A* does not do anything that *B* had not expected *A* to do (conventionally) and *A* is not hurting herself.

Resentment: If *A* acts *unfairly* in the sense above (that is, unpredictably), *B* will feel resentment toward *A*.

Resentment aversion: Players who cause resentment in other people's minds forfeit utility. Thus, utility maximising agents are resentment-averse.

So Sugden (2000), in effect, defines 'fairness' as conformity with the evolved status quo. Anything that frustrates others' expectations is deemed unfair, goes against the grain of their expectations, and is the cause of negative psychological utility. People, in this view, are driven by the psychological desire to avoid the disapproval that comes from frustrating others' expectations.

14

Granted that we all experience a certain dissonance from causing resentment in others, it is still unlikely to be the only primitive urge. It seems to us that humans equally have a *subversive tendency* (that is, our tendency to want to subvert others' expectations of us); otherwise, it is likely to be difficult to explain how people ever consciously escape the status quo. Formally, we might define our proclivity to subverting others' beliefs as follows: if *A* expects *B* to expect *A* to perform *X* and yet *A* chooses some other action, *Y*, for the purposes of causing resentment

in *B* (through frustrating *B*'s expectations about *A*), then *A* is being purposefully subversive.

Clearly, subversion is a disequilibrium phenomenon as it implies that *A*'s higher order beliefs about *B* are out of alignment. By contrast, Sugden's *resentment-aversion hypothesis* is an equilibrium notion since it relies on common knowledge that *B* has good reason to form the empirical expectation that *A* will do *X* rather than *Y*.

The subversion-proclivity hypothesis (definition)

Conformism: In equilibrium, person *A* is a *conformist* as long as *A* does not do anything that *B* had not expected *A* to do (conventionally) and *A* is not hurting herself.

Subversion: If *A* acts *contrary* to *B*'s expectations (that is, unpredictably), *B* will think of her as subversive and will feel a combination of *resentment* and *admiration* toward *A*.

Subversion proclivity: Players who gain net utility from causing in others this combination of *resentment* and *admiration* are characterised by *subversion-proclivity*.

'The main constituents of a satisfied life appear to be... two: tranquillity and excitement,' wrote John Stuart Mill. The reader will notice immediately the dependence of *subversion-proclivity* on the prior evolution of Sugden's *resentment-aversion* as well as the tension between the two. When these two tendencies are played out in historical time, and in the context of the simultaneous evolution of behaviour and motivation, the result is a never-ending cycle between periods of stability (during which some convention is established in accordance with the *resentment-aversion hypothesis* – RAH) and subsequent periods of flux (during which older conventions are being disestablished, in accordance with our *subversion-proclivity hypothesis* – SPH). It is interesting to recall that this conflict of primitive (though not irrational) urges, was the foundation of the critique of subgame perfection in

Chapters 3

and

4

; i.e. that games like the *Centipede* (or Rubinstein's, 1982, bargaining game), are indeterminate due to the irrepressible tension between an equilibrium and a subversive logic.

To support SPH, we need two things. First, we need to link SPH to some primitive human psychological trait (as Sugden did with his RAH). Secondly, we need a plausible story as to how the proclivity to subvert and frustrate others' expectations by subversive actions has been reinforced through the evolutionary process. With regard to the former, it seems to us that we are often torn between seeking others' approval though conforming with their expectations and wanting to impress (others as well as ourselves) through uncommon behaviour; behaviour that helps us 'stick out'.

The tensions between these two urges arises because, often, the most effective way of getting noticed is to frustrate others' expectations about us and (at least initially) cause them to resent us. Indeed, causing resentment in others (at least initially) may be a prerequisite for the success of our strategy to impress and get noticed. This is a

fascinating aspect of our confused, and at once majestic, nature that we often admire persons for precisely the same reasons for which we also resent them. The question now is: what is its social function and how did it come about?

In the case of the *resentment-minimisation* psychological trait, Sugden's neo-Humean (evolutionary) explanation is clear: conformity generates regularities which help populations reduce the chances of costly conflict. However, by the same token, we can explain evolutionarily the reinforcement of the subversive trait if we can show that a periodic purge of established conventions increases a community's fitness.

As is well known in the literature, a well-established convention may well be 'inferior' compared to alternative ones (e.g. the inefficiency of QWERTY). Indeed, a discriminatory convention which reduced conflict effectively in the past (by arbitrarily advantaging one subpopulation over another) may have exceeded its use-by date (e.g. as a result of technological change). Therefore, communities benefit from a capacity to undermine (and thus test for the evolutionary stability of) what Sugden refers to as *normative expectations equilibria*. If that capacity is related to the subversive trait in us all, one can argue that the evolutionary process reinforces *at once* two contradictory traits of human nature: *resentment-aversion* and *subversion-proclivity*.

I mentioned above the possibility that, in stratified societies, a person belonging to a disadvantaged group can gain substantial kudos from subverting the established discriminating convention, at least within her own group. To make this point, however, I need to re-draft our SPH in terms consistent with evolution in more than one dimension.

The one-dimensional subversion-proclivity hypothesis (ODSPH)

Suppose there is a population P and some convention C which has evolved earlier one-dimensionally (recall

[Section 6.2](#)

). By definition, in interactions of a given kind between members of P , C recommends to each person i the same action X (as opposed to Y). If some person j chooses Y , then this choice will engender a degree of resentment in the person she has interacted with and even among the rest of the population (assuming common knowledge of j 's behaviour). Finally, suppose there has been a history H of continual choices in accordance with C by all members of P . Then (and this is the hypothesis) if j chooses Y , she will secure a degree of notoriety, admiration etc. proportional (a) to H and (b) to the degree of resentment caused by her choice of Y . Thus, as long as persons within P have a taste for notoriety, admiration etc. (however small), there exists some H which will trigger subversion.

The two-dimensional subversion-proclivity hypothesis (TDSPH)

The difference with **ODSPH** above is that (the previously evolved) convention C is two-dimensional and discriminatory. That is, it segregates (on the basis of some

arbitrary feature) population P (conventionally) between two subpopulations (P_1 and P_2) and gives different instructions to $i \in P_1$ and to $j \in P_2$: It directs i to play, in a meeting with j , X and j to play Y . Suppose that i 's utility is such that $U_i(i \text{ plays } X, j \text{ plays } Y) > U_i(i \text{ plays } X, j \text{ plays } X) = U_i(i \text{ plays } X, j \text{ plays } Y)$. Finally, if there is a history H of continuous adherence to C by members of both subpopulations, then j 's choice of X (rather than Y) in violation of C will lend her some psychological utility from 'sticking out,' notoriety etc. which is proportional (a) to H , and (b) to the resentment caused among members of subpopulation P_1 .

So, the obvious parallel with evolutionary biology is to think of SPH as equivalent to mutations testing the stability of the established evolutionary equilibrium C . Will individual subversion succeed in undermining C ? It depends on its capacity to spread by infecting others. One might speculate that in the one-dimensional case, the chances of subversion are limited. Each subversive move will be a tiny drop lost in a sea of conformity. Nevertheless, even under those circumstances, conventions are disestablished and customs change when the bandwagon of a new norm is ready to roll. The world of fashion is one area that comes to mind.

The multi-dimensional case is, of course, far more interesting. Norms of honour among gentlemen are functional to norms of excluding women from the benefits of equality. Since convention C segregates P into subpopulations, each with its own behavioural pattern and normative/calculative expectations about the other, the success of subversive moves will clearly depend on whether j 's subversion will give rise to collective acts of subversion by members of subpopulation P_2 . To the extent that such 'collective spirit' is functional to the interests of subpopulation P_2 , the emergence of correlated deviations from C (by members of P_2) are likely to be associated with other 'bonding' practices within P_2 , e.g. greater reluctance to succumb to the norm of adhering to mutual defection in the *prisoner's dilemma*, or sub-population-specific lifestyle choices *vis-à-vis* music, fashion etc. To give a celebrated example, the defiance of a sole middle aged black woman riding on a segregated bus in the American South would have gone unnoticed in the 1960s had there followed no coalition of black men and women who turned her subversive act into a campaign.

More grandly, it is tempting to claim that SPH lies behind behaviour which helps society discover not only new *ways* to play given games but of new *games* to play as well. In

Chapter 12

below, I shall lament evolutionary game theory's reliance on fixed payoff structures. It will be the reason I shall pronounce it insufficiently ... evolutionary. In this chapter, however, the idea that payoffs are contingent on beliefs allows us to imagine a genuinely evolutionary theory of society.

15

From this perspective, one might expect a theoretical account involving a mixture of (often opposing) social forces constantly equilibrating and subverting the evolving 'system'. As a result, one would expect to find periods of continuity which are interrupted by severe discontinuities not only in the behaviour but, importantly in the structure of the social interaction (i.e. of the dominant games). At the level of beliefs, history makes itself felt in the never-ending establishment

and (subsequent) subversion of normative belief equilibria. At the level of the cultural, the primitive appeal of subversion manifests itself in the best works of drama and literature.

8.4 Epilogue: shared praxes, shared meanings

This chapter parted ways with conventional game theory in one important respect: it linked beliefs *directly* to desires. The result was, I wish to argue here, reminiscent of a potentially interesting distinction between game theory's rules of the game, which are

regulative, with Wittgenstein's rules of language games, which are *constitutive*.

It is indeed possible to interpret the norms of Rabin (1993) and Sugden (2000) as akin to the rules of a Wittgensteinian language game. This interpretation seems plausible because, in this chapter, norms are no longer simple regulative devices (as they were in previous chapters). They do a lot more than simply help satisfy pre-existing preferences (as they might be doing in a Humean or neo-Humean account). In fact, they help *constitute* the players' actual preferences. Interpreting the rules is quite different to subscribing to them.

The analogy is helpful because it ties in with the change to the existence of symbolic properties associated with *action*, namely with their *meaning*. On Wittgenstein's view, the attribution of shared meaning to words in a language cannot come from some shared experience of either the external world or our inner feelings. *Shared meanings depend on shared practices*. This is a controversial claim because it depends in part on the impossibility of holding a private language. Nevertheless, it makes us social from the outset, with language marking this fact, as does the existence of norms above, rather than either being a derivative from some version of exchange between pre-social individuals.

Throughout this book, I have made clear my objection to the reduction of human *reasonableness* to the assumption of *instrumental rationality*. The conclusion here, concerning the impossibility of knowing what one wants *outside a web of shared practices*, is grist to that mill. One of the great gifts of game theory to the social sciences is that it has caused some thoughtful economists to question the assumption of instrumental rationality. The present chapter is based on results that emanated from such scepticism within the economics profession.

This re-think has caused me to return to the very first questions an intelligent novice might ask before even beginning to grapple with neoclassical game theory: *What is a game? How is it constituted?* Beyond saying that a game is a situation in which the outcome for one participant depends jointly on the actions of all, the answer must address the crucial issue of the players' motivation. Neoclassicists deal with this issue concisely and without much discussion: players have preordained preferences over the range of outcomes and they act in a manner that satisfies these preferences.

By contrast, this chapter has shown that motivation is much more complex than this. Inspired by Wittgenstein, it comes in the form of a suggestion that is dynamite under the neoclassicist's lazy philosophical premises: Players' perception of

their preferences is ill-defined before the game is played. More formally, what is instrumentally rational to do is not well-defined unless one appeals to the prevailing norms of behaviour. But, if the prevailing norms of behaviour are the result of playing the game, then the game and our ways of playing it are codetermined as part of a joint evolutionary process.

This may seem a little strange in the context of a superficial reading of neoclassical textbooks on, say, the one-shot *Prisoners' Dilemma*. In that game, game theorists proclaim, the demands of instrumental rationality seem plain for all to see: *Defect!* But, in reply, I would protest the presumption of payoffs which have fallen *as if* out of thin air, unvarnished by social experience. Just like dogs and humans, so humans and ... games evolved side-by-side over millennia. The norms that govern our behaviour also govern our interpretation of the events unfolding around us. A social setting requires *interpretation* before we know what we want and how much we value different outcomes. But if the same norms that govern our behaviour are also implicated in those interpretations, how can we claim that motives are prior to games?

In conclusion, the study of psychological games has clarified the game theorist's dilemma: she may continue to pursue game theory's Holy Grail of 'closing' game theoretical explanations without 'outside' assistance. Or she may admit that

Indeterminacy has won the day anyway, and use analyses like those offered in this chapter in order to understand the limits of neoclassicism. If I am right, game theory will keep tilting at the windmills of *Indeterminacy* until it goes out of fashion as the futility of this task becomes evident. That would be a shame. For game theory has a lot to offer, as this chapter has demonstrated. It is a powerful tool with which to explore liberal individualism's limits and the difficulties of conjuring up satisfying social explanations. To go beyond this requires a change. Rather than 'solving' insoluble strategic interactions, or thoughtlessly applying existing 'solutions', the point is to figure out what games we play, how these came about and, perhaps, how we ought to change them.

VERDICT: This chapter is perhaps the pinnacle of the dance of the meta-axioms, this book's symbolic theme. It has investigated a challenge to neoclassical orthodoxy that sprang out of neoclassicism's own underbelly. It began when some enlightened game theorists dared ask an impertinent question: *What happens when we allow people to evaluate their options not just on the basis of projected outcomes but also in terms of what they think that other participants expect of them (either in predictive or normative terms)?* This is a fascinating question that some game theorists have pursued imaginatively, enthusiastically and intelligently.

Alas, as is always the case, their 'challenge' led them straight into the *Wall of Indeterminacy* (recall the diagram in

Chapter 1

). And what a species of indeterminacy it turned out to be! Indeterminacy on steroids. It was not just that their models succumbed to outcome indeterminacy, as they invariably do, and required a strong dose of neoclassicism's third meta-axiom to be 'salvaged'. No, it was much, much worse than that: They required a mega-dose of the third meta-axiom just in order to allow for the players' utility functions (i.e. preferences over

outcomes) to be well-defined. Even then, after the third meta-axiom (without a smidgeon of logical support) was deployed to 'fix' players' preferences, the outcomes remained indeterminate – perhaps the only example in this book where even the third meta-axiom cannot help a neoclassical model produce determinate outcomes.

Naturally, these neoclassical forays (dating to the early 1990s) into the more interesting aspects of human motivation died out and have been conveniently allowed to slip into oblivion. Once the repercussions of these models were understood by journal editors and more generally by the powers-that-be who rule over neoclassicism's institutions, theorists like Matthew Rabin were silenced on these matters and forced to move to other realms of more anodyne (from neoclassicism's perspective) inquiry.

It was not, of course, that anyone explicitly censored people like Rabin. No, what happened was that the indeterminacy of their models meant that they were returned by the prestigious journals marked 'unpublishable' and that colleagues in departmental seminars would refrain from putting on display any show of excitement during their presentations. Thus, the dance of the meta-axioms had, once again, performed its miracle of maintaining and reproducing neoclassicism's most peculiar failure.

Appendix 8.1: What came first, capitalism or the profit motive?

The idea that games cannot be well-defined outside the realm of the social norms that emerge when people play them sounds like a typical catch-22 problem. Of course, economics is replete with such chicken-and-egg problems, just as the rest of social theory is immersed in them. Indeed, we can see that in the most basic debates, from the very outset of political economics. Adam Smith, for example, suggested that the division of labour in society was dependent upon man's 'propensity to barter, truck and exchange one thing for another.' This phrase was later to yield the concept of *homo*

economicus whose clones populate all economics and game theory texts. Polanyi (1945) famously challenged Smith's view that there is something *natural* in people that turns them into merchants when the opportunity arises. According to Polanyi, Smith misread the past (by recognising potential merchants in the serfs, Lords and artisans of pre-capitalist societies). But, he added, '...[i]n retrospect it can be said that no misreading of the past ever proved more prophetic of the future...' (Polanyi, 1945, pp. 50–1). Polanyi's own view was that the newfangled motives (i.e. the propensity to barter etc. for profit) emerged at the same time, and for the first time, as genuinely new social games (that is, market societies, or capitalism) were being formed on the ruins of the feudal era:

The outstanding discovery of recent historical and anthropological research is that man's economy, as a rule, is submerged in his social relationships. He does not act so as to safeguard his individual interest in the possession of material goods; he acts so as to safeguard his social standing, his social

claims, his social assets. He values material goods only in so far as they serve this end. (Polanyi, 1945, p. 53)

The above view is consistent with this chapter's analysis of the psychological aspects of payoffs. The pursuit of social standing gives rise to different motivations, depending on the prevailing norms. Without knowing the norms, it is impossible to know their motivation. The two evolve, and bring new patterns to the fore, simultaneously. Neither the game nor the motivation comes first.

Karl Marx has often been disparaged for not grounding his theories of capitalism on the individual. His reasons can be seen more clearly in the light of the present discussion: For if the individual is not prior to capitalism, nor vice versa, what is the scope of any theory (e.g. methodological individualism) which takes the individual's motives as givens and only then tries to explain society against the background of these given motives? Marx's chosen solution was to deal with individuals theoretically:

...only in so far as they are the personifications of economic categories, embodiments of particular class relations and class interests. My stand point, from which the evolution of the economic formation is viewed as a process of natural history, can less than any other make the individual responsible for relations whose creature he socially remains, however much he may subjectively raise himself above them.

(Marx, Preface to the first German Edition of *Das Kapital*)

Notes

- 1 Note that c is a best reply to c when $3a > 4(a - b)$ or $b/a > 1/4$. When this inequality holds, c is a best reply to c while d remains a best reply to d . In this sense, we have two Nash equilibria in pure strategies and a strategic structure identical to that of the *stag hunt*.
- 2 Note that this threshold is given as $q^* = b/(\alpha + \beta)$. It rises with the material benefits from the pernicious c and falls with the sum of the psychological parameters α and β . Clearly, as sum $\alpha + \beta$ increases (and/or b falls), the threshold value of q^* declines and the probability that $q > q^*$ increases. In short, $p = Pr(q > q^*)$.
- 3 The average payoffs below are computed thus: We have already discovered that strategy c will be adopted if $d > 0$ or $q < b/(\alpha + \beta) = q^*$. Jill will be indifferent between the two and will therefore choose a mixed strategy if and only if $d = 0$ or $q = b/(\alpha + \beta) = q^*$. In that case, in equilibrium, $p = q^*$ and Jill's average payoffs will equal $(1 - q^*)(b - \alpha q^*) + q^* \beta q^*$. Substituting $q^* = b/(\alpha + \beta) = q^*$ into this expression yields Jill's average payoffs under psychological equilibrium (iii).
- 4 The following classification of Rabin's (1993) assumptions, along with their labels, are not to be found in the original paper; they reflect my own interpretation of his paper.
- 5 Combined with *reciprocity*, this definition of neutrality implies the following: When Jill expects that Jack is being neutral toward her, she feels no urge to be either kind or nasty back. That is, unless neutrality is reciprocated with neutrality, psychological utility is forfeited ($\Psi < 0$).
- 6 Note how we have gone from one to two equilibria even in the humble prisoner's dilemma which, hitherto, featured a unique dominant strategy per player.

7 Clearly, as μ rises psychological utility matters more to this person than material rewards and vice versa.

8 Note that a defecting B is making no sacrifice regardless of whether he expects A defect or to cooperate.

9 Since when AbBbA:c A's relative gain is greater than when AbBbA:d, while B's sacrifice by cooperating is the same in both cases

10 To give an example, consider the game below. Suppose that A expects B to play strategy c . If she responds with c too, she is clearly being kind. The reason is that she is playing a non-Nash strategy (the cooperative c as opposed to her aggressive best reply h) in order to aid B. Suppose, however, that she replied with another non-Nash strategy: d . B would again benefit (albeit less) from her non-Nash behaviour at her expense: he would receive payoff 0 as opposed to -1 . Expression (8.8), however, determines that d is not really an act of kindness, even though B benefits at A's expense; it is, rather, an act of folly on A's part. If she wanted to make a sacrifice on his behalf, she should have chosen c . In this manner, both B and A would benefit most from reciprocated kindness. For this reason, function f_A takes the value zero when A responds to c with d : it simply rules out inefficient behaviour (i.e. a kind of foolishness) as a case of rational kindness.

		<i>B</i>		
		<i>h</i>	<i>d</i>	<i>c</i>
<i>A</i>	<i>h</i>	$-2, -2$	$2, 0$	$4, -1$
	<i>d</i>	$0, 2$	$1, 1$	$0, 0$
	<i>c</i>	$-1, 4$	$0, 0$	$3, 3$

11 This is in sharp contrast with Rabin whose rather crude specification insists that A's choice of h in response to B's d is always tantamount to an unkind act.

12 The reason is that if B chooses d because he anticipated d from A, he is not hurting A at a cost to himself. Thus he means her no ill and deserves no nastiness from her. It is this thought that renders dd a mutual kindness-neutral equilibrium. On the other hand, if one cooperates with a defector one is being hugely kind (a kindness value of 2 is reported) toward a kindness-neutral person.

13 The reader who has not grasped this point yet may see it clearly by observing that the kindness functions f_A and f_B (see Table 8.3) are determined by the players' entitlements (as in Rabin) which are in turn determined (and this is our innovation) by the player's second-order beliefs.

14 Notice that fairness was defined differently in the previous sections: Something more was demanded from A before we could proclaim her action 'fair': a degree of sacrifice (however small) when compared to what A could have got away with; a sacrifice that would lead to a benefit, or a loss, from someone who made a similar sacrifice to aid or pain us. In short, and unlike Sugden (2000a), to be fair one needed to pay a price.

15 Of course, our account has left most interesting psychological categories out of the analysis. The features of shame and guilt, which often guide human behaviour, are two examples of motivation which is not subject to our control. However, this chapter does contain interesting pointers for some of the absent psychological categories. For instance, the neo-Humean attitude toward shame and guilt is not hard to imagine: deliverances of illusions bestowed upon us through social evolution. There is nothing *objectively* wrong, they might argue, with littering the streets or killing our mothers. It is just that society has become more stable and better able to reproduce itself when we all live under the fantasy that it is wrong to such things. As for shame and guilt, they are the psychological mechanisms which provide the requisite motiation at the level of our 'souls'.

9 The social foundations of corruption

On the indeterminate power of what others think

9.1 Prologue

9.1.1 Background briefing

This chapter attempts to demonstrate the usefulness of the previous chapter's analysis, despite neoclassicism's abandonment of this rich vein that, ironically, neoclassicists themselves had unearthed. Just as

[Chapter 5](#)

demonstrated that the indeterminacy we unveiled in

[Chapter 4](#)

did not impede useful insights from emerging, similarly here I shall attempt to show that psychological game theory has the capacity to help us understand phenomena that are real, important and hitherto ill-understood because of, rather than despite, the indeterminacy it generates. In short, the contribution of chapters such as

[Chapter 4](#)

and the present one is to point out that neoclassicism, because of its obsession with 'closed' models, has a penchant for jettisoning into the abyss insights that it itself has contributed to social science. Just like a monopolist who is keen to destroy part of the surplus in order to maintain his own monopoly profits, neoclassicism destroys part of its own intellectual surplus so as to keep non-neoclassical theorists at bay.

The theme of this chapter, to which elements of last chapter's psychological game theory will be applied, is the joint evolution of (a) corruption and (b) public engagement in politics. Theoretically speaking, this is accomplished by means of a model combining psychological game theory with evolutionary game theory. Its contribution is to demonstrate that, while power corrupts and corruption undermines the legitimacy of power, the prospects for social and economic development may depend crucially on the evolution of an appropriate web of expectations, rather than on a powerful coercive mechanism that forces corruption underground. The theoretical results emphasise the context-specificity of corruption, explain resistance-to-corruption as a response to preferences inhabiting the ill-defined space between the walls separating one citizen from an 'other', and links the evolution of corruption to the evolution of public-spiritedness and the reach of participatory politics.

¹

9.1.2 The rest of the chapter

Corruption is usually modelled in a Hobbesian manner: officials have an interest to further their private goals, without any in-built interest in the means by which they will succeed, and are deterred solely by the threat of punishment from some type of Leviathan (e.g. the legal machinery of the State, administrative checks and balances, loss of reputation).

²

Similarly, participation in the public sphere is also modelled in a Hobbesian manner, usually assuming cynical citizens who only participate in institution-building or democratic politics if there is 'something in it for them'.

This chapter looks at another source of resistance to corruption and at a different motive for participating in the public sphere: the *psychological impact of others' expectations*. The hypothesis, explored analytically in

[Chapter 8](#)

, is that our utility from corrupt acts is influenced, independently of their consequences (e.g. whether we get away with it or not), by the level of propriety others expect of us. Similarly, our utility from participating in the public sphere depends on whether others expect us (or not) to contribute to the pursuit of collective objectives. In

short, the means we employ to achieve our ends are a source of non-consequential utility whose magnitude depends on our second-order beliefs (i.e. our beliefs regarding what others expect of us).

The basic idea relates well to the thought that coercion, however essential, is a rather inefficient tool for preventing corrupt practices, as the costs involved in keeping corruption at bay exclusively through a punishment mechanism are enormous. Similarly, public spiritedness would be a very rare bird if agents did not derive a sense of satisfaction from investing in the public sphere; a sense contingent on other people expecting of them such contributions.

David Hume (1888), writing in 1740, knew this only too well: conventions, he tells us, are best preserved when agents begin to 'believe' in their preservation; when they acquire a normative dimension. In short, societies that are largely corruption-free are the ones in which officials would not act corruptly even when no one is watching over them. And those in which we find a high level of participation in public affairs are the ones in which citizens have internalised into their preference set, under considerable peer pressure, the urge to invest in public goods and institutions.

On the other hand, a disdain toward corruption, or an urge to be a decent citizen, are neither born out of nothing nor maintained simply by one's occupation of the high moral ground. More often than not, it is sustained by the expectations of others. In this sense, it is a counterpart of solidarity – another term that addresses the cracks between the Hobbesian and the Kantian extremities (see

[Chapter 7](#)

).

When others expect you to act 'properly', this very thought is often enough to make you *want* to act properly (even if you could get away with acting corruptly or meanly). To put it differently, when others hold you in high esteem and expect propriety or public spiritedness from you, there are serious psychological costs involved in acting corruptly or selfishly. Moreover, these costs play a significant part in impeding corrupt behaviour and defeating the privatisation of the self.

Of course, for the bonds of others' expectations to take hold, society must first reach some 'equilibrium' in which most people anticipate 'proper' behaviour from their officials, bureaucrats or citizens. The rest of the chapter offers an evolutionary model featuring a number of different equilibria. In some of these equilibria, the

weight of others' expectations create bonds that impede corruption and promote participation; in others they do not.

9.2 Corruption, apathy and the fragility of collective agency: a static interaction

9.2.1 Introduction

Power corrupts and corruption erodes the legitimacy of power. Societies that are plagued with corruption find it hard to shake citizens out of their apathy so that they can participate in the creation of practices that broaden political accountability (see Emerson, 2006). Thus, the prospects of effective collective agency and democratic politics depend heavily on whether corruption and apathy are 'states' which society's evolutionary dynamics favours or selects against (see Emerson, 2002; Mauro, 1995). Our task here is to offer a model of the evolution of corruption in conjunction with a depiction of the dynamics of citizen participation in political life. It will be a simple model, borrowing both from *evolutionary game theory* (see Weibull, 1995) and *psychological game theory* (see the previous chapter). Despite its simplicity, it brings out some important insights regarding the critical importance of second-order beliefs in helping society steer a course away from an equilibrium in which apathy and corruption slow down economic and societal development (see also Dahlberg and Mork, 2006;

Verbrugge, 2006).

The model is based on two interactions: First, a game between bureaucrats, in which each player must choose her level of corruption. Corrupt practices bear private material benefits but come at a variety of costs. Secondly, an interaction between citizens who decide whether or not to participate in the formation of policy and institutions (e.g. to become politically active).

In this model, there is no chance that a corrupt official will be caught and punished. In other words, the Hobbesian enforcement mechanism is taken out so as to explore the limits to corruption that can be effected purely by costs related to status and personal embarrassment. The costs of corrupt behaviour modelled below fall under two types: (a) the loss of reputation that the political system, or the 'class' of bureaucrats, experience *as a whole*, and (b) the private psychological costs from acting corruptly, which are taken as proportional to the bureaucrat's second-order belief that she is upstanding (i.e. acting corruptly costs her more psychological utility the higher her estimate of the citizen's expectation that she is honest). Naturally, the dividing line between pecuniary and non-pecuniary costs is thin: the loss of reputation as the public cottons on to the officials' corruption can be readily translated into reduced future income (e.g. if the government's tenure is somehow related to their reputation) or simply kept at the level of the psychological effects of some loss of social stature.

As I show below, the bureaucrats' game spawns a number of evolutionary and psychological equilibria. The model concludes that the public's expectations play a significant role in determining the equilibrium level of corruption. The interesting point here is that, unlike other models lacking a psychological component,

in this model citizens' expectations guide not only what bureaucrats will do but, importantly, their *preferences*, viz. the final outcome as well.

The citizens' game featured further below is a standard free rider problem, augmented also with a psychological-cum-political component. The free rider aspect of the interaction stems from the fact that the personal cost of participating in the democratic process is high relative to the individual's capacity to affect, through her own (isolated) political efforts, the 'public good'. (Think of the tedious meetings, the opportunity cost of demonstrating, writing letters to editors etc. in a society where a crushing majority abstain from such practices.) The novel 'psychological-cum-political' component introduced below is based on a simple thought: The private (psychological) benefit of participation in political activity is inversely proportional to the person's estimate of the degree of corruption by bureaucrats, politicians etc. That is, the greater one's expectation that the bureaucratic/political class is steeped in corruption, the less rewarding political activity becomes at the grassroots level.

This 'component' provides the primary linkage between the two games (the interaction between bureaucrats and that between citizens). A secondary link is introduced such that the bureaucrats' expected material benefits are a decreasing function of citizen participation. The simple idea here is that corruption succeeds more, on average, in societies populated by apathetic citizens. The interesting effects of psychological variables in one interaction spill over, with sometimes surprising results, to the other. As evolutionary pressures weed out corruption, political participation flourishes. And vice versa. In some special cases corruption may survive high levels of participation but never vice versa.

9.2.2 An interaction between bureaucrats

This section presents the interaction between bureaucrats. Corruption is, to them, a means of gaining material benefits. Usually, models of corruption assume that the only impediment to acting corruptly is the threat of punishment. In this model, such a threat is absent: its purpose is to explore the possibility of corruption-free equilibria in the absence of surveillance and administrative punishment mechanisms. Once this

theoretical task is achieved, it is straightforward to introduce a Hobbesian dimension (e.g. a probability of being caught when acting corruptly as well as a series of punishments, depending on the level of corruption).

Central to the following model is the idea that the private interests of a particular bureaucrat may clash with their collective (or regime) interests. This raises the important issue of whether, and to what extent, bureaucrats shall coordinate their individually rational actions in a bid to achieve their common and private goals.

Let us now turn to the crucial question: In the absence of the fear that they will be caught and punished, what stops them from acting corruptly in this model? The answer turns on the introduction of (a) psychological effects on corrupt bureaucrats depending on what the public expected of them *ex ante*, and (b) *ex post* reputational effects that measure the importance of social status. These two ‘complications’ are introduced by means of two key assumptions:

ASSUMPTION 9.1 *Bureaucrats suffer psychological disutility when acting corruptly toward a citizen who expected higher standards of honesty from them. By contrast, whenever the public expects high levels of corruption from some bureaucrat, the latter is released from this psychological impediment and suffers no disutility from acting corruptly.*

ASSUMPTION 9.2 *The bureaucrat’s utility contains a component which is proportional to average opinion (among citizens) regarding the honesty of people like herself (i.e. of bureaucrats).*

Let $c_i \in [0, 1]$ denote bureaucrat i ’s chosen level of corruption;

$p_i = \Pr(1 - c_i)$ be the probability with which bureaucrat i will select level of honesty $1 - c_i$ (or, equivalently, level of corruption c_i);

$p'_i = E_{\text{public}}(p_i)$ be the public’s average estimate of p_i ; and

$q_i = E_{\text{bureaucrat } i}(p'_i)$ be B ’s estimate of p'_i ; i.e. q_i is B ’s second-order belief regarding the probability with which she will be honest.

Now define U_i as the utility function of bureaucrat $i = 1, \dots, N$. In accordance with assumptions (1) and (2), U_i comprises three components:

(a) $M(c_i)$ is the bureaucrat’s *pecuniary utility*, or utility from material gains, which is *ceteris paribus* an increasing function of her ‘corruption’ level: $M'(c_i) > 0$

(b) $\Psi(q, c_i)$ is the bureaucrat’s *non-pecuniary utility*, or disutility (see Assumption 9.1), from her corrupt actions when she thinks that the public expect from her a level of propriety or honesty equal to q . [Reflecting Assumption (

9.1

) above, the higher the value of q the lower $\Psi(q, c_i)$ becomes as c_i rises.]

(c) $L[(\sum c_i)/N]$ represents *social status utility*; that is, utility from the aggregate reputation or status of bureaucrats. The assumption here is that the bureaucrats’ collective status is inversely proportional to aggregate corruption. Clearly, $L'(\cdot) < 0$

Let us posit an additive utility function for the typical bureaucrat:

$$U_i = M(c_i) + \Psi(q, c_i) + L\left(\left(\sum c_i\right)/N\right) \quad (9.1)$$

To explore the trade-offs facing i I impute the simplest functions that capture the spirit of the above:

$$M(c_i) = \text{constant} + \beta c_i; \quad \Psi(q, c_i) = \text{constant} - \gamma q c_i;$$

$$L(q) = -\alpha \left(\sum c_i\right)/N \quad (9.2)$$

(where α, β and γ are positive parameters)

Thus, the typical bureaucrat’s utility function is given by:

$$U_i = \text{constant} + (\beta - \gamma q)c_i - \alpha \left(\sum c_i\right)/N \quad (9.3)$$

Note that parameter β represents the marginal rate of *pecuniary utility* from corrupt

practice; parameter γ is the marginal non-pecuniary disutility from corrupt behaviour *given the public's expectation that bureaucrats are honest*; and parameter α captures the rate at which overall utility is lost when the bureaucrats collective reputation for honesty declines (or, equivalently, their collective reputation for corruption rises).

From (

9.3

) it is clear that, as long as $\beta < \alpha$, bureaucrats would prefer a state in which all of them are honest to one of pervasive corruption. For if $c_i = 1 \forall i$ then the public expects maximum corruption ($q = 0$) and $U_i = \text{constant} + \beta - \alpha$. On the other hand, wholesale propriety means that the public expect no corruption from bureaucrats ($p' = 1$), the latter are utterly incorruptible ($p = 1$), they know that the public thinks so ($q = 1$) and each one of them collects payoff $U_i = \text{constant}$. Thus, if they had a choice between wholesale corruption and wholesale propriety, all bureaucrats would opt for the latter as long as $\alpha > \beta$.

However, even in this case corruption may emerge as the only equilibrium outcome if the bureaucrats are caught in the clutches of a type of prisoner's dilemma. For instance, suppose that, indeed, $\alpha > \beta$. Even though our N bureaucrats would suffer if they all acted corruptly (in comparison to their utility from across-the-board propriety), each will have a dominant strategy of acting corruptly as long as $\partial U_i / \partial c_i > 0$. Noting that $\partial U_i / \partial c_i = \beta - \gamma q + \alpha / N$, it transpires that the bureaucrats are caught in an N -person prisoner's dilemma (or free-rider problem) as long as

$$q < q^* = (\beta N - \alpha) / \gamma N \quad (9.4)$$

Intuition: Widespread corruption will occur once the bureaucrats' reputation for propriety falls below a certain threshold (q^*) *even when their collective interest suffers as a result* (i.e. when $\alpha > \beta$).

Equilibria: In equilibrium, since bureaucrats are identical, they adopt the same level of corruption, say c ; the public have accurate expectations of average honesty and, therefore, expect c from each bureaucrat ($p' = 1 - c$), each bureaucrat knows that the public's estimation of average corruption is $q = p' = 1 - c$ and, consequently, each bureaucrat's utility level is given by (5) below:

$$U_i = \text{constant} + (\beta - \alpha)c - \gamma(1 - c)c \quad (9.5)$$

Equilibrium Type I: $c = 1$, $q = 1 - c = 0$. All bureaucrats act corruptly and the public anticipates no propriety on their part. From (5), $U_i = \text{constant} + (\beta - \alpha)$

Equilibrium Type II: $c = 0$, $q = 1 - c = 1$. All bureaucrats act properly and honestly and the public anticipates no corruption on their part. From (5), $U_i = \text{constant}$

Equilibrium Type III: $q = q^* = (\beta N - \alpha) / \gamma N = 1 - c$ and $c = [(\gamma - \beta)N + \alpha] / \gamma N$. A proportion q^* of bureaucrats act with propriety, this is anticipated by the public, and the average bureaucrat receives a utility payoff of $U_i = \text{constant} + \{(\beta - 2\alpha) [(\gamma - \beta)N + \alpha]\} / \gamma N$

Note. From (4), we know that $Pr(\text{Type I equilibrium}) = Pr(c = 1) = Pr[q < q^*]$; $Pr(\text{Type II equilibrium}) = Pr(c = 0) = Pr[q > q^*]$; and $Pr(\text{Type III equilibrium}) = Pr(c = [(\gamma - \beta)N + \alpha] / \gamma N) = Pr[q = q^*]$ where $q^* = (\beta N - \alpha) / \gamma N$. In equilibrium, however, q can take only three different values: 0, 1 and q^* . From these observations we deduce the following necessary conditions: For a *Type I* equilibrium, the necessary condition is $q^* > 0$ (otherwise q can never be less than q^*). For a *Type II* equilibrium, the necessary condition is $q^* < 1$ (otherwise q can never exceed q^*). And for a *Type III* equilibrium, the necessary condition is that $[(\gamma - \beta)N + \alpha] / \gamma N$ falls within the range [0, 1]. These necessary conditions (for equilibrium *Types I, II* and *III* respectively) can be simplified as follows. *Type I:* $N > \alpha / \beta$; *Type II:* $[\alpha / (\beta - \gamma)] > N$; *Type III:* $[\alpha / (\beta - \gamma)] > N > \alpha / \beta$. From these necessary conditions, it transpires that as N rises, *Type I* equilibrium becomes more prevalent.

Case 1 – $\alpha > \beta$. In this case, bureaucrats dislike the prospect of wholesale corruption.

A *propriety equilibrium* is preferable to a *corruption equilibrium*. Although the marginal pecuniary utility resulting from corruption (a) may be high, the marginal non-pecuniary utility losses from increases in the bureaucrats' own perception of how corrupt the public expect their type (or regime) to be (b) are even higher. Thus, bureaucrats would prefer a *Type II* to a *Type I* equilibrium.

However, this does not mean that they will *necessarily* refrain from corruption. For if $q < q^* = (\beta N - \alpha)/\gamma N$ [see (9.4) above], each has an incentive to act corruptly (even though they all prefer that all remain incorruptible!).

Let us consider a numerical example:

Example $\alpha = 10$, $\beta = 1$, $\gamma = 3$, $N = 100$

In this case, $q^* = 0.3$ and there are three potential equilibria:

Equilibrium Type I: $c = 1$, $q = 1 - c = 0$ All bureaucrats act corruptly and the public anticipates no honesty on their part. From (

9.5

), $U_i = \text{constant} - 9$

Equilibrium Type II: $c = 0$, $q = 1 - c = 1$ All bureaucrats act properly and the public anticipates no corruption on their part. From (5), $U_i = \text{constant}$

Equilibrium Type III: $q = q^* = 0.3 = 1 - c$ and $c = 0.7$. That is, either 70 per cent of the bureaucrats are corrupt or each bureaucrat chooses a level of private corruption equal to 0.7 (on the 0 to 1 scale); this is anticipated by the public; and the average bureaucrat receives a utility payoff of $U_i = \text{constant} - 13.3$.

Case 2 – $\alpha < \beta$. Bureaucrats prefer a *corruption-equilibrium* (*Type I*) to an *honesty-equilibrium* (*Type II*). However, again this does not automatically mean that the former will prevail. Interestingly, the prisoner's dilemma's logic cuts both ways. For example, if $q > q^*$ the average bureaucrat will be better off (due to high psychological rewards from her interaction with citizens) to remain honest even if she preferred a *corruption-equilibrium*. In this case, the bureaucrat wishes that the public expects the worst from her. Such low public expectations would liberate her from the internal restraint [courtesy of her non-pecuniary utility $\Psi(\cdot)$]. To illustrate, consider the following numerical example:

Example $\alpha = 4$, $\beta = 10$, $\gamma = 20$, $N = 100$

Clearly, *Type I* equilibrium (a *corruption-equilibrium*) is preferred by each bureaucrat to the alternatives. Nevertheless, we note that $q^* = 0.498$ and, thus, yet again there are three potential equilibria:

Equilibrium Type I: $c = 1$, $q = 1 - c = 0$ All bureaucrats act corruptly and the public anticipates no honesty on their part. From (9.5), $U_i = \text{constant} + 5$

Equilibrium Type II: $c = 0$, $q = 1 - c = 1$ All bureaucrats act honestly and the public anticipates no corruption on their part. From (9.5), $U_i = \text{constant}$

Equilibrium Type III: $q = q^* = 0.498 = 1 - c$ and $c = 0.5025$. That is, either about 50 per cent of the bureaucrats are corrupt or each bureaucrat chooses a level of private corruption equal to 0.5025 (on the 0 to 1 scale); this is anticipated by the public; and the average bureaucrat receives a utility payoff of $U_i = \text{constant} + 1.004$.

Summing up, this section has demonstrated two things: Independently of whether the bureaucrats (or politicians) prefer wholesale corruption or all around propriety, they are susceptible to a prisoner's dilemma logic capable of subverting their collective interest and one which depends on the public's expectations of the bureaucrats' demeanour. We examined two such cases of unintended consequences: One in which bureaucrats want to see all their colleagues behave properly and honestly. In this case, whether they shall be caught in the trap of the prisoner's dilemma (and end up all corrupt and relatively dissatisfied) depends on what the public expects of them. The second case was one in which bureaucrats had no compunction: they would be quite

happy to be part of a comprehensively corrupt regime. Interestingly, even in this case, the public's expectations can put them in a prisoner's dilemma situation which renders overall propriety the game's unique equilibrium.

The gist here is that what matters most in determining the pervasiveness of corruption is not so much the bureaucrats' own preferences but the public's perception of them. If the public expects high standards of behaviour from its bureaucracy, it will have them regardless of the latter's collective preference. In this sense, during a period when the public (or electorate) is losing its confidence in a certain regime, government, administration etc. this loss of esteem may indeed 'liberate'

bureaucrats and encourage them to become even more disagreeable.

Section 9.3

makes this insight more explicitly obvious by subjecting the above analysis to the evolutionary approach. But before we come to it, we need to model the interaction between citizens which determines the level of their participation in public affairs.

9.2.3 Participation in public affairs and institutions: an interaction between M citizens

Just as our N bureaucrats select their level of corruption, $M(>N)$ citizens select the level $\pi_j \in (0, 1)$, $j = 1, \dots, M$, at which they wish to participate in the public sphere, the democratic process etc. In a manner reflecting the bureaucrats' earlier concern for what the public expected of them (viz. their corruption levels), the citizens' decision to participate in the public sphere is influenced by what others expect of them. To capture this idea, we let

$r_j = Pr(\pi_j)$ be the probability with which j will select participation level π_j

$r'_j = E_{\text{public}}(r_j)$ be the public's average estimate of π_j ; and

$s_i = E_j(r'_j)$ be j 's estimate of r'_j i.e. s_i is j 's second-order belief regarding the probability with which she will participate in public affairs.

The citizen's utility function W_j comprises three components:

$$W_j = b \left(\frac{\sum_{g=1}^M \pi_g}{M} \right) - k \times \pi_j + \rho s \times \pi_j \quad (9.6)$$

The first component captures the utility benefits to person j from living in a society in which citizens participate actively in the fashioning of the public agenda, institutions, policy etc. It depends, naturally, on average participation $(\sum \pi/M)$ and some multiplication factor $b(>1)$ describing the way in which each unit of j 's participation translates into private satisfaction from this public good. For example, if all M citizens participate fully (i.e. if $\pi_j = 1 \forall j$) then everyone derives utility equal to $b(>1)$ from the public sphere.

The second component arrests the private cost $k(<b)$ of each unit of chosen participation. For example, attending meetings, writing letters to the local newspaper, standing for office etc. come at a considerable personal cost; in our case, k units of lost utility per unit of π .

The third component is similar to the non-pecuniary $\Psi(\cdot)$ subutility component of the bureaucrats' utility function in (1): Parameter $\rho(>0)$ helps ameliorate the private costs of having selected participation rate π_j in proportion to j 's impression of the extent (s) to which *others* expect her to participate. When $s = 0$, j thinks that no one expects her to participate in the public sphere and, thus, this third component vanishes. However, the more she thinks others expect her to participate (i.e. the higher her second-order belief s) the more she enjoys participating

independently of her participation's contribution to the common good (i.e. independently of b). To put it differently, the higher s the lower j 's private cost from participating.

The above lead inexorably to a model of the decision to participate that has all the

hallmarks of a public good or free rider problem. In large communities the free rider aspects of this decision rear their head the moment we acknowledge the possibility that one person's participation in political affairs has an infinitesimal effect on the common good (b/M) but, at the same time, it brings substantial costs to the said individual.

Analytically, if $\partial W_j / \partial \pi_j = b/M - k + \rho s < 0$, citizens land in the trap of the prisoner's dilemma and thus idly watch by as participation in the public sphere withers. This outcome is, of course, not deterred in the least by the thought that all citizens would prefer a full participation equilibrium to a complete apathy one (note how individual utility W_j equals $b - k + \rho > 0$ when $\pi_j = 1 \forall j$ and zero when $\pi_j = 0 \forall j$).

9.3 An evolutionary version of the two interactions

9.3.1 A two-dimensional replicator dynamic

So far, our interaction between bureaucrats and the one between citizens were both of a static type. Agents chose their control variables (bureaucrats selected their c_i 's and citizens their π_j 's) only once and in isolation. This section offers a full evolutionary version of both interactions.

Consider a large population of identical bureaucrats interacting randomly and repeatedly with a large population of citizens. Furthermore, all citizens interact with one another in the participation game. To render the model evolutionary, we need two mechanisms: A replicator dynamic per interaction plus a mutations' mechanism. The replicator dynamics are modelled simply as follows:

In the bureaucrats' interaction, we found that that as long as $\partial U_i / c_i > 0$, or as long as $q < q^* = (\beta N - \alpha) / \gamma N$ [recall (

9.4

)], the average corruption levels rises and the bureaucrats' second-order beliefs regarding their level of propriety (q) diminishes – and vice versa. To illustrate, we suppose that, as time goes by, whenever the net gains from corruption exceed zero, the frequency $(1 - \rho)$ of honest acts shrinks. Consequently, so will the public's prediction of honesty on the part of bureaucrats $(1 - \rho')$. And as the public loses confidence in the bureaucrats, the latter will work this out and their estimate of ρ' , q , will fall. Note the main presumption here: Bureaucrats' behaviour adapts gradually as bureaucrats switch to the strategy/action with the positive net gains.

Meanwhile, a similar process is unfolding in the interaction among citizens that determines participation in the public sphere and, thus, s (where s is defined as the citizens' second-order belief concerning their participation). As long as $\partial W_j / \pi_j < 0$, or $b/M - k + \rho s < 0$, citizens gradually abandon the public sphere, participation dwindles, the public expects that its members will be recoiling into their own private worlds and thus fewer and fewer people will expect that others will expect of them high levels of participation.

In short, the two *replicator dynamics* are as follows:

$$\frac{\partial U_i}{c_i} > 0 \quad \text{or} \quad q < q^* = (\beta N - \alpha) / \gamma N \Rightarrow q \text{ falls} \quad (9.7a)$$

$$\frac{\partial W_j}{\pi_j} < 0, \quad \text{or} \quad s > s^* = (kM - b) / \rho M \Rightarrow s \text{ falls} \quad (9.7b)$$

Mutations mechanism: In accordance with evolutionary theory (see Weibull, 1995), and in order to ensure the stability of evolutionary equilibria, we assume that at all times there are random 'deviations' whereby some bureaucrat acts honestly (corruptly), even at a time when corruption (honesty) pays better on average, or that some citizen gets interested in politics against the calculus of her utilities. This 'mutation mechanism' allows us to examine the evolutionary stability of equilibria.

Figures 9.1

and

9.2

capture the evolutionary path of the two interactions.

Figure 9.3

combines the two separate processes into a single diagram. From these it is clear that, the evolutionary process is driven fully by citizens' expectations. Starting with

Figure 9.1

, we see that the standards of the bureaucracy depend on the public's initial expectations, and the way the bureaucrats perceive these. If the public starts off with low expectations regarding the bureaucrats' propriety (and the latter know this), they precipitate low standards amongst bureaucrats who, in a never-ending circle, end up confirming the low opinion of them held by citizens (evolutionary equilibrium $q = 0$). And vice versa: 'Great expectations' on behalf of the public lead bureaucrats (whether they like it or not) to *want* to act with propriety!

Figure 9.2

tells a similar story regarding the level of the public's own participation in public life. If citizens expect that their fellow citizens expect them to participate at a rate greater than a certain threshold (s^*), then participation burgeons and the full participation evolutionary equilibrium results ($s = 1$). Otherwise, apathy rules.

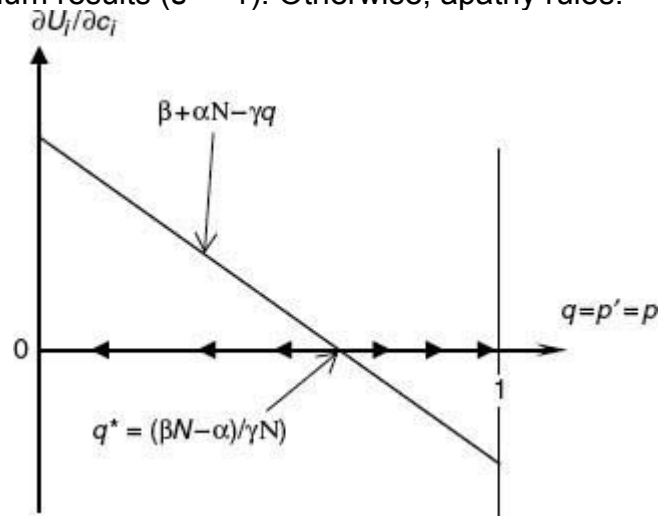


Figure 9.1

Two evolutionary equilibria in the bureaucrats' interaction: $q = 0$ or 1 .

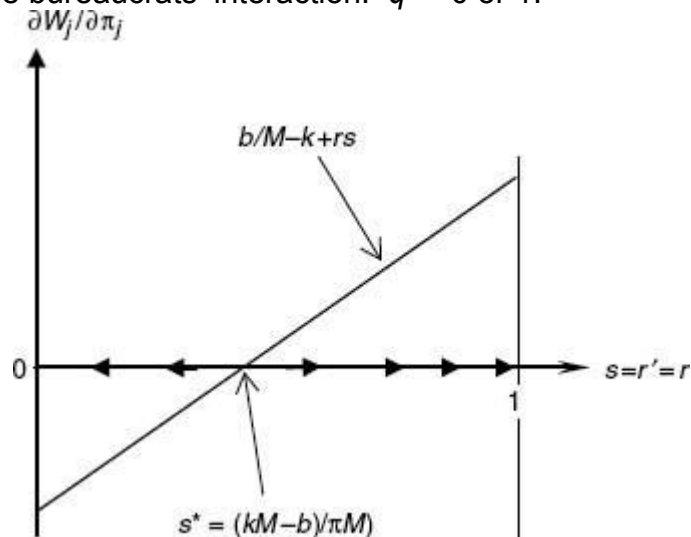


Figure 9.2

Two evolutionary equilibria in the citizen participation interaction: $s = 0$ or 1 .

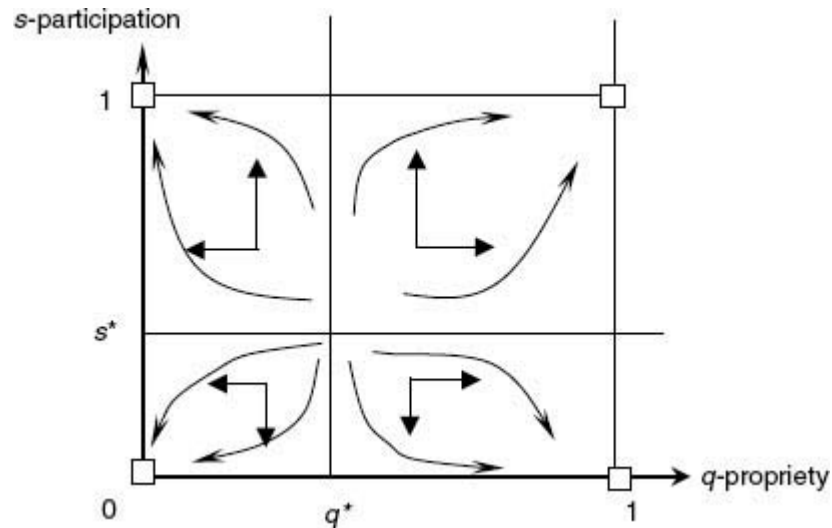


Figure 9.3

Four evolutionary equilibria when the two interactions are *not* linked: $(q, s) = (0, 0), (0, 1), (1, 0), (1, 1)$.

Lastly,

Figure 9.3

puts together these separate evolutionary processes in the context of a single two-dimensional diagram. The conclusion here is that, as long as the two evolutionary interactions remain un-linked, society may end up in any of the evolutionary equilibria at the four corners of the diagram: (1) Corruption plus apathy (bottom left), (2) Propriety plus apathy (bottom right), (3) Participation and Corruption (top left), and (4) Participation and Propriety. In the next section I show that some of these potential equilibria are eliminated if the two interactions are linked.

9.3.2 Linking the two interactions: or, how citizen participation limits the evolutionary possibilities of corruption, and vice versa

The simple idea in this section is that the time paths of corruption among bureaucrats and of citizen participation are somehow linked. In particular, I shall presume that low standards of propriety by officials turn citizens to apathy. The opposite is, of course, also possible: 'Great' expectations on behalf of citizens energise bureaucrats against corruption and, in turn, keep the fire of political activism burning.

The proposed linkage takes a simple analytical form. Parameters b and β [see equations (9.6)

and (

9.3

) respectively] now become variables. The former is the multiplication factor which influences the private enjoyment of citizens from the existing degree of political vibrancy in one's society. I assume that this is a positive function of the equilibrium level of propriety by bureaucrats (q). The latter, β , is the bureaucrat's marginal pecuniary utility gain from corrupt behaviour. I shall now assume that β is a decreasing function of the equilibrium level of citizen participation in public life (s). In simpler terms, as corruption (or its perception) gathers pace, private 'joy' from making a contribution to the public sphere declines, thus accentuating further the free-rider aspects of the intra-citizenry interaction. At the same time, when citizens participate increasingly in public life, the life of corrupt bureaucrats becomes more 'difficult' and their marginal pecuniary benefits fall. In short,

$b = b(q)$, with $db/dq > 0$ and $\beta = \beta(s)$, with $d\beta/ds < 0$

To keep the analysis as uncomplicated as possible, I shall posit linear functions as follows:

$$\text{For each bureaucrat: } b = b_0 + \lambda q \quad (9.8a)$$

$$\text{For each citizen: } \beta = \beta_0 - \mu s \quad (9.8b)$$

(where λ and μ are constant rates of change, and b_0 , β_0 are the base line values of these 'parameters').

Introducing (

8a

) and (

8b

) into the twin replicator dynamics of

equations (7a)

and (

7b

), we derive new replicator dynamics in (

9a

) and (

9b

):

$$q < q^* - \mu s / \gamma \text{ or } s < \Lambda(q) = \gamma / \mu (q^* - q) \Rightarrow q \text{ falls} \quad (9.9a)$$

$$s > \Xi(q) = s^* - \lambda q / \rho M \Rightarrow s \text{ falls} \quad (9.9b)$$

Geometrically,

equations (9.9a)

and (

9.9b

) give rise to eight different configurations – see

Figures 9.4

–

9.11

. The main finding, in juxtaposition to

Figure 9.3

, is that the introduction of feedback between the two interactions restricts considerably the number of evolutionary equilibria. In particular, we see that in five of the eight possible cases (see

Figures 9.3

–

9.5

,

9.9

and

9.11

) only two of the original four evolutionary equilibria remain: Either society will be corrupt and its citizenry apathetic ($s = q = 0$) or corruption will be routed out by a society whose citizens participate fully in public affairs ($s = q = 1$). Even in the remaining three cases, one of the equilibria in

Figure 9.3

has dropped out. In two cases (see

Figures 9.7

and

9.11

) it is no longer evolutionarily possible to have full participation while bureaucrats are corrupt, while in the remaining case (see

Figure 9.8

) wholesale corruption is impossible to achieve in the presence of full citizen participation.

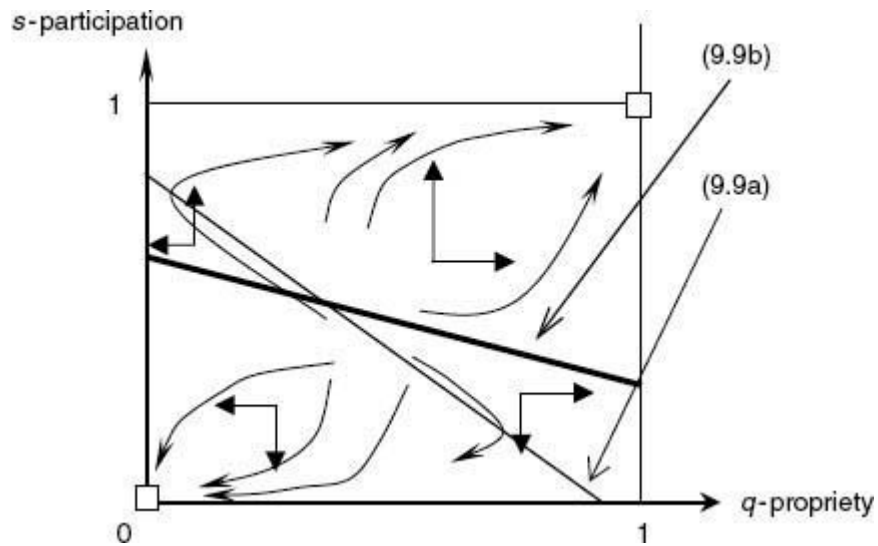


Figure 9.4

Two evolutionary equilibria: $(q, s) = (0, 0), (1, 1)$.

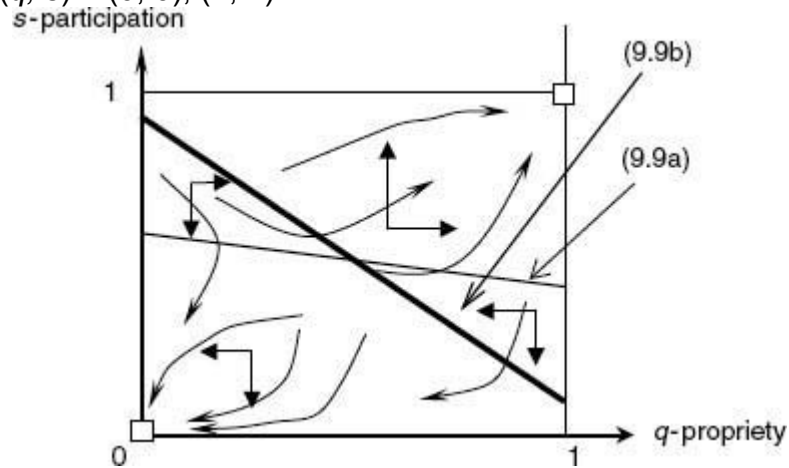


Figure 9.5

Two evolutionary equilibria: $(q, s) = (0, 0), (1, 1)$.

In short, making the pecuniary rewards from corrupt practices a decreasing function of citizen participation, while assuming that the latter declines with the incidence of corruption, restricts nicely the range of potential equilibria. In effect, such linkages give rise to evolutionary paths that impose a stricter alliance between participatory politics and resistance to corruption.

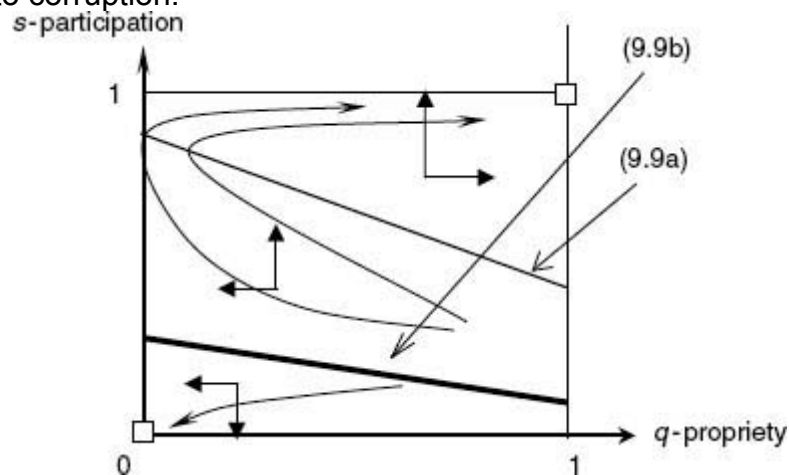


Figure 9.6

Two evolutionary equilibria: $(q, s) = (0, 0), (1, 1)$.

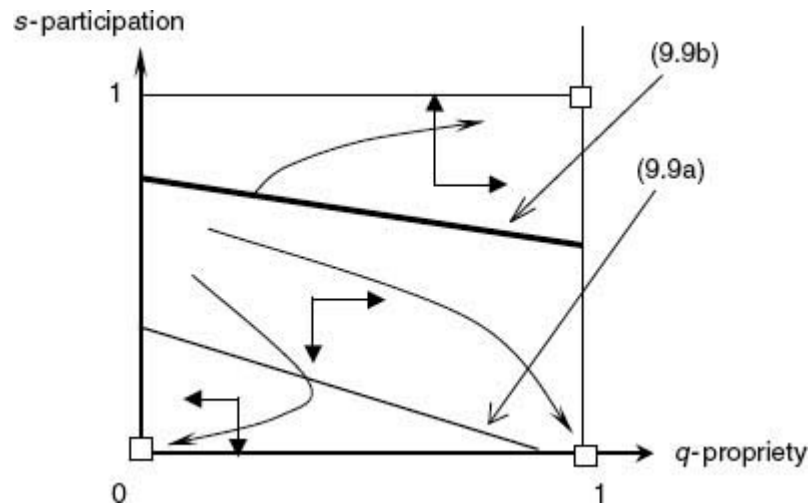


Figure 9.7

Three evolutionary equilibria: $(q, s) = (0, 0), (0, 1), (1, 1)$.

9.4 Epilogue

While power corrupts, and corruption undermines the legitimacy of power, the prospects for social and economic development may depend crucially on the evolution of an appropriate web of expectations (first- and second-order), rather than on a powerful coercive mechanism that forces corruption underground. This point was driven home by the preceding analysis, based on a simple model which:

- (a) shows how corruption may not be inevitable *even if no Hobbesian enforcement mechanism is in place*
- (b) emphasises the context-specificity of corruption (in ways that resemble the work of anthropologists, e.g. Harrison, 2006)
- (c) explains resistance-to-corruption as a response to preferences inhabiting the ill-defined space between the walls separating one person from an 'other' *and*
- (d) links the evolution of corruption with the evolution of public spiritedness and the reach of participatory politics.

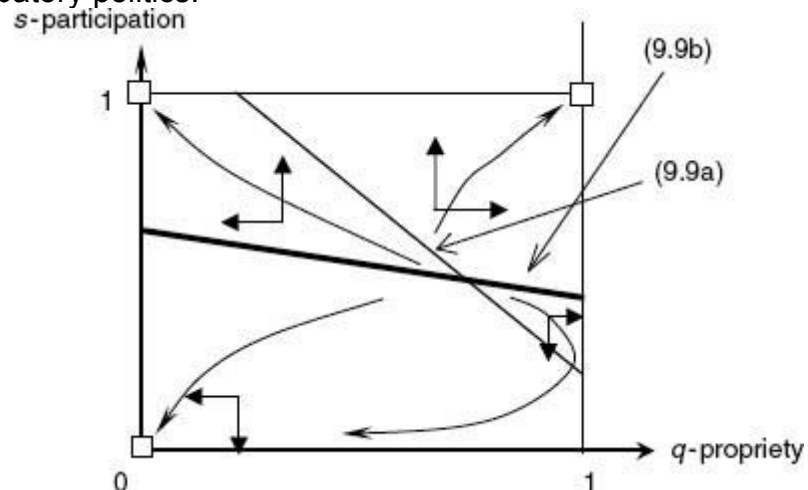


Figure 9.8

Three evolutionary equilibria: $(q, s) = (0, 0), (1, 0), (1, 1)$.

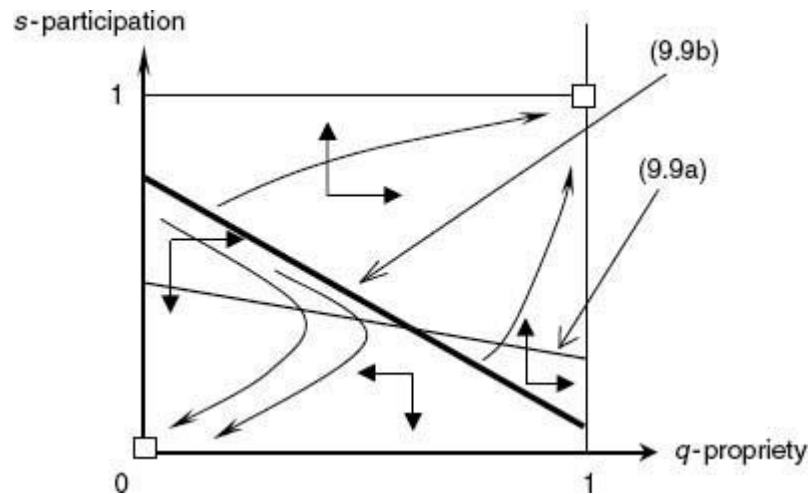


Figure 9.9

Two evolutionary equilibria: $(q, s) = (0, 0), (1, 1)$.

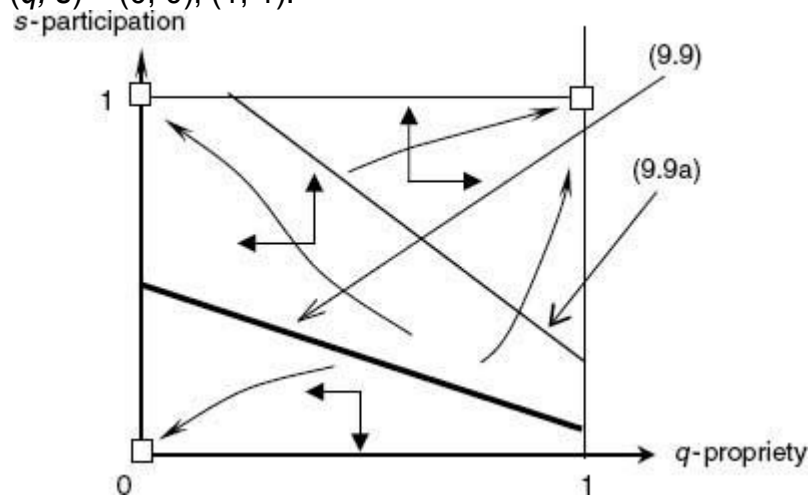


Figure 9.10

Three evolutionary equilibria: $(q, s) = (0, 0), (1, 0), (1, 1)$.

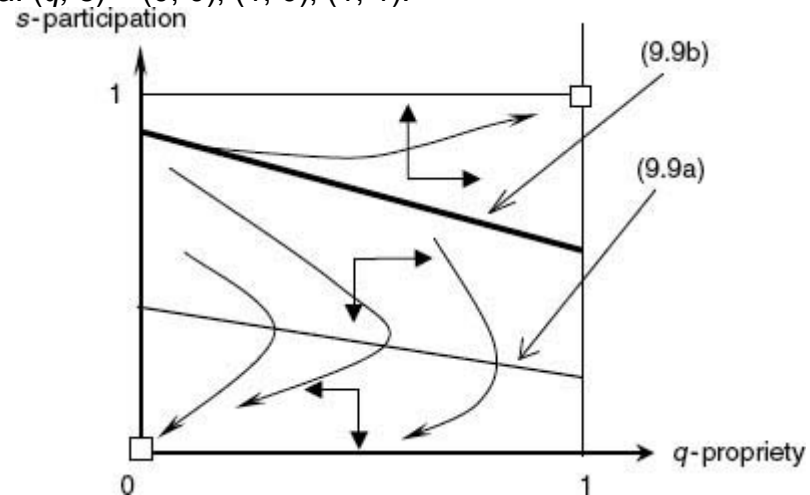


Figure 9.11

Two evolutionary equilibria: $(q, s) = (0, 0), (1, 1)$.

The broader question, which the chapter did not address, concerns the links between the evolutionary-*cum*-psychological process under discussion and the evolution of the society's material/industrial foundation. However, in accordance with this book's theme, and in conjunction with preceding chapters, this chapter *did* manage to open a window onto a most peculiar form of corruption: that which neoclassical economics infuses into academic discourse and with which it hollows up the moral centre of young academics drawn into its bosom. I wrote above that 'power

corrupts, and corruption undermines the legitimacy of power'. This may be so in almost every realm of public life, except of course Economics Departments where, through the dance of the meta-axioms, intellectual fraud and corruption of the mind extends and reinforces the legitimacy and power of a form of economics which is constitutionally illegitimate, logically incoherent and whose discursive power is detrimental to civilised society.

Notes

1

This chapter is based on Varoufakis (2006).

2

See Ales and Tella (1996), Auriol (2006), Bliss and Tella (1997), Saha and Thamby (2006), Schleifer and Vishny (1993).

3 Note that parameter γ is multiplied with q . This means that the negative impact of the public's perception of the corrupt bureaucrat's utility is an increasing function of the latter's perception of the bureaucrats' average reputation for honesty among members of the public.

4

Note that, in equilibrium, the psychological effects from personal corruption drop out and parameter γ does not affect the bureaucrat's preferences between equilibria of *Types I* and *II*. This happens because, when all bureaucrats are corrupt (*Type I*), the public expects no honesty and thus the individual bureaucrat suffers no psychological loss from being corrupt since no one expects her to be otherwise. And when they are all honest (*Type II*), no psychological losses are suffered (by definition).

10 Evolving morals in the laboratory

The roots of distributional justice principles in indeterminacy

10.1 Prologue

10.1.1 Background briefing

As an undergraduate student I vividly recall being told, by venerable academic economists, that economics is an empirical science. The whole ideology of 'positive' economics was based on the view that economists are a species of empiricist who have no commitment to their theories and who use them only in order to formulate testable hypotheses. Thus, the claim that economics is all about working out how the social world works, without any prejudice regarding how it 'ought' to function.

This romantic notion – that theories are just initial hypotheses that we put in the path of reality, allowing them to be crushed by the force of evidence as we learn more about social reality – is quite appealing to the inquiring mind. Until, that is, one came across ... econometrics. To cut a long story short, it took me a couple of years, while completing a PhD thesis in micro-econometrics, to realise an awful truth: econometrics has nothing to do with the pursuit of truth by means of sacrificing expendable theories on the altar of empirical evidence. Precisely because there never existed even the remotest possibility of a one-to-one relationship between a theory and a reduced form (i.e. the empirical function that was to be tested against the data), and because viciously competing theories corresponded to the *same* reduced form: the great theoretical disputes could never be adjudicated by the 'facts'.

This realisation, early on in my career, led me to the firm conclusion that economics was a hopeless enterprise. On the one hand, the neoclassical models that dominated the discipline were being 'closed' by hidden assumptions which guaranteed their irrelevance to really-existing capitalism. On the other hand, econometrics would, and could, neither expose this irrelevance nor help guide us to at least a modicum of understanding about the deeper causes of economic and social phenomena. Was there no hope of enlightenment?

At that time, around 1984, I moved from the University of Essex (where I was writing my thesis) to the University of East Anglia (as a young, underpaid lecturer). East Anglia had just recruited Bob Sugden, a thoughtful economist who had already established a career as an experimenter. Bob's presence energised a

few faculty members to take laboratory experiments seriously, bringing into the debate, concerning their uses and prospects, Shaun Hargreaves-Heap – my future co-author – and philosopher Martin Hollis. In that context, I began thinking about the possibility that laboratory experiments might provide an escape route from the theoretical and empirical straitjacket of neoclassical economics and econometrics.

To give a flavour of the promise of experiments, consider 'expected utility theory,' which Bob Sugden and others had taken to task experimentally. Theoretically, expected utility theory is a coherent model of instrumentally rational choices. Because it can be used as a foundation on which to build all sorts of fancy neoclassical models (from consumption and portfolio theory to game theory), it has a very special place in the neoclassical toolbox and, of course, curriculum. To challenge it philosophically is a futile project that neoclassicists would not even honour with a passing glance. However, given neoclassicism's adherence to the ideology of positivism, showing that under well-designed laboratory conditions (which ought to give expected utility theory its best chance to succeed predictively) people systematically and predictably violate expected utility theory's basic tenets, is another matter. At the very least, neoclassicists are discomfited by such evidence and feel compelled to put up a defence; to take the bearers of such 'bad' news seriously.

In summary, I was drawn to laboratory experiments in a bid to challenge empirically

neoclassicism's certainties. This chapter, and the next, report on two major experiments that occupied me, and Shaun Hargreaves-Heap, for almost a decade. As you will see, dear reader, experimenting in a laboratory is not just great fun. It also leads to surprising results that one could never have imagined a priori. That these results force neoclassical economists to engage, however briefly, with one's objections to their theoretical presumptions, is an added bonus. Alas, let me state clearly that any hope that neoclassicism may be forced, by the force of empirical evidence coming from a laboratory, to do anything other than pause for a few minutes (before returning to the reproduction and re-capitulation of its arid and toxic models) is bound to crash on the shoals of reality. ...

10.1.2 The rest of this chapter

This chapter is dedicated to an experiment (published in Varoufakis, 1997) the purpose of which was to demonstrate that (a) conventions may well emerge which discriminate against certain groups of people, (b) the selection of persons to be discriminated against has nothing to do with their personal attributes (e.g. intelligence, risk aversion), and (c) these discriminatory conventions infect agents with particular beliefs that help the conventions become more powerful and stable.

The idea behind this experiment can be traced to ancient arguments and puzzles. Here I narrate the said experiment with a story from Thucydides' *Peloponnesian War*, which reminds us that moralising, from the standpoint of strategic weakness, has always been a last resort strategy. According to Thucydides, the ancient Melians presented the Athenian generals with a splendid example when in a particularly tight corner. In our Western philosophical tradition moral rhetoric is often couched in the form of reasons for action either external to preference and desire

(e.g., Kant) or internal to the agent's calculus of desire (e.g., Hume, Gauthier, neoclassical economics). A third tradition dismisses such rhetoric as the last recourse of the weak (e.g., Aristotle, Nietzsche) whereas a fourth calls for an examination of the social context (e.g., Socrates, Marx, Wittgenstein, Habermas). This chapter relies on an experimental study in order to throw some empirical light on these debates and offers a surprising twist to the interpretation of the Melians' plea.

Section 10.2

presents the Melians' argument, explains the experiment that I used to test some hypotheses related to the Melians' claims, and discusses the rationality and moral content of strategic rhetoric. Then

Section 10.3

presents and analyses the empirical evidence that our experiment generated.

Section 10.4

puts this debate in a broader philosophical context. Finally,

Section 10.5

offers the chapter's epilogue.

10.2 The Melians' plea and a relevant experiment

10.2.1 Moral principle or clever tactic?

Morality has been hailed variously as a product of enlightened selfishness, the greatest proof of our autonomy, a social construct, an elaborate illusion; the list goes on. Regardless of the perspective, its relation with strategy and justice has a long lineage. Thucydides reports that, in the course of its geopolitical struggle against Sparta, Athens dispatched a fleet with the specific order that the independently minded island-state of Melos be subdued or razed to the ground. In the dialogue entered into by representatives of the two sides, following the arrival of the Athenian assault troops, the interplay between moral principles and strategic concerns underscored the rhetoric.

In an opening speech, anticipating Aristotle's infamous pronouncement that '[t]he weaker are always anxious for justice and equality. The strong pay heed to neither'

(*Politics*, 1 s1318), the Athenians demanded Melos' surrender. After all, they decried, on the one hand the principles of justice, encompassed in human reason, hinge on the equal capacity to compel, yet on the other hand, the strong actually do what is possible and the weak suffer what they must.

(Thucydides, *The History of the Peloponnesian War*, Book 5, s89)¹

The Melians, at that point, played their only card. They demanded that they be allowed to remain neutral and free for Athens' own sake:

Then in our view (since you force us to base our arguments on self-interest, rather than on what is proper) it is useful that you should not destroy a principle that is to the general good – namely that those who find themselves in the clutches of misfortune should be justly and properly treated, and should be allowed to thrive beyond the limits set by the precise calculation of their power. And this is a principle which does not affect you less, since your own

fall would be visited by the most terrible vengeance, watched by the whole world.

(Thucydides, *The History of the Peloponnesian War*, Book 5, s90)

Necessity invented a splendid, and highly prophetic, argument: When in the dominant position do to others what you would like to be done to you when weak. If your behaviour is unconstrained by such a *principle*, you will live to regret it.

Years later, these words resonated in Athenian ears as the Spartans scaled the walls of Piraeus, intent on destruction. Of course, Thucydides does not elaborate on the precise philosophical content of the Melians' argument. Were they envisaging principled behaviour as the solution to the calculus of long-term Athenian preferences, or were they canvassing a universalisable principle to be activated by a pro-active reason (e.g., Kant's advice: 'Act only on that maxim which you can at the same time will that it should become a universal law')? To help unravel this ancient mystery, let us pay a visit to a laboratory where modern subjects are observed while interacting. Their behaviour, this chapter contends, may hold clues to the Melians' strategy.

10.2.2 The experiment

The experiment in question inadvertently produced some interesting insights concerning the Melians' plea. It involved several hundred volunteers who played three versions of a simple game. To get a handle on how the games were played, consider the first version of the game as described below (two further versions will be introduced later).

Suppose you are sitting in front of a computer terminal which assigns you role R: you are asked to choose a row strategy from the set (1, 2, 3). At the same time, someone else (whom you do not know and you cannot see) is choosing among the set of column strategies 1, 2 or 3. [I shall refer to the player who chooses among the rows as an R-player (or R) and the one who chooses between column strategies as a C-player (or C).] The payoff matrix in

Table 10.1

translates such a pair of choices into payoffs; for example, if you choose 2 and your partner chooses 2, then you win nothing and your opponent wins 5.

Table 10.1

is the first of three versions of the same type of game (see

Table 10.3

for the other two versions). Each version was played four times by each subject, who alternated between the two roles: if they chose among the rows in one round,

they chose among the columns in the next and among the rows in the following round, etc.

Table 10.1

Version 1 of the game played by subjects in the laboratory

	C1	C2	C3
--	----	----	----

R1	5, 0	-1, -1	-1, 10
R2	-1, -1	0, 5	-1, -2
R3	-1, 10	-2, -1	6, 6

The basic structure of the game is this: If both players choose their third strategy (R3 or C3) they collect the same payoff of 6. No other payoff can be better for a player, *except if the other player receives a negative payoff* [as in the case of outcomes (R1, C3), (R3, C1)]. If players do not choose 3 and 3, then one player will get (at best) 5 while the other will receive zero [outcomes (R1, C1) or (R2, C2)]. At worst they will both suffer a negative payoff [(R1, C2), (R2, C1)]. Thus, the combination of strategies (R3, C3) corresponds to the only *mutually* beneficial outcome, which I shall refer to as the *cooperative outcome*.

The problem with the *cooperative outcome* (R3, C3) is that its prospects are seriously undermined by a logic very similar to that which undermines cooperation in the case of the standard prisoner's dilemma. Consider again the version of the game above: Cooperation (i.e., choosing one's third strategy: R3 or C3) is threatened by two thoughts: First, if you want to achieve the cooperative outcome (R3, C3), you will fear that your opponent may anticipate this and choose his or her first strategy in order to collect the largest available payoff of 10; in which case you are left with a negative payoff (-1). Second, even if you trust the other person not to 'cheat' in this manner, *you* are tempted to 'cheat' since selecting your first strategy, in response to your opponent's third, will reward you with payoff 10. These two thoughts reinforce each other and sabotage the chances of the cooperative outcome.

In short, no payoff-maximising player will chose R3 or C3 since they can do better by choosing one of their other strategies. Suspicion that one's opponent might not recognise this, will tend to encourage R-players to choose 1 and C-players to choose C1. Increasingly, the third strategies will be abandoned.

10.2.3 Acts which under-utilise one's strategic advantage

Would it make sense for someone to choose the cooperative strategy in our game? No, is the instrumental answer, for which neoclassical economists are primarily responsible. Regardless of whether you are an R or a C player, if you suspect that your opposite number will select her or his third strategy, your payoff-maximising response is to select your first strategy and in so doing collect the highest possible payoff of 10. In the language of game theory, the cooperative strategies are *dominated* (and thus instrumentally irrational choices). Furthermore, the experimental design (that is, the fact that the games are played anonymously and, in each round, each subject is paired against a different opponent) removes any tangible reason to expect that subjects will care about anything other than short-term (i.e., round by round) payoff maximisation. Thus the instrumental expectation that, in this interaction, rational subjects will not cooperate.

Does it make sense to defy this logic and cooperate? In one sense, urging a player to play cooperatively in this simple interaction is similar to the Melians urging the Athenians to spare them. Although a lot less is at stake in the laboratory, in both cases agents are asked to disregard their strategic calculations. Before

observing our subjects' behaviour, let us recount some well-rehearsed explanations of why people may set aside their strategic possibilities and behave in a cooperative, quasi-moral, manner similar to that advocated by the Melians (see

[Table 10.2](#)

).

From an instrumental viewpoint, there are three explanations worth considering. One is that they have failed to recognise what it is in their interest to do; that they have

misread the situation and did not realise that cooperating is a dominated strategy. In this case [(1a) in

Table 10.2

] agents are expected to pay more attention to the strategic structure of the interaction (and thus cooperate less) the greater the stakes (i.e. the larger the numbers in the payoff matrix) and the more experienced they are.

The second instrumental explanation, (1b), requires a shared future of mutually recognisable agents so that a good reputation for cooperativeness may yield long-term benefits. Then the instrumental agent may bite her tongue and cooperate in the short run in order to enjoy a string of 6-payoffs in the medium run (what game theorists refer to as trigger strategies or tit-for-tat). The reason why this does not qualify as moral (or principled) behaviour is that when we reach the last interaction, our agent will immediately shuffle off her reputation and abandon the pretence of being a cooperative soul. Indeed, an interesting line of thought applies here according to which the finiteness of most interactions wrecks the chances of such enlightened selfishness, regardless of the size of the long term gains. (For a discussion see Pettit and Sugden (1989); Hollis (1991); Varoufakis (1993) – or

Chapter 4

of this book.)

The final instrumental explanation (1c) takes us to Hume's (1888) *Treatise* where he argues that the agent's morality is to be found in her passions, inclinations or preferences (as opposed to her reason). The moral agent shows some natural sympathy to the preferences of others and can rationally act on them in a manner which cannot be explained if we only take into consideration her personal gains. This would mean that the payoffs in the matrices above do not reflect the true preferences of individuals. For instance, players may derive an additional 5 'psychic' units from the cooperative outcome because they value a cooperative outcome *per se*. Thus, it would be instrumentally rational to cooperate.

Table 10.2

Six explanations for acts which seemingly defy the agents' strategic interests

1. *Instrumental*
 - a. Execution errors (e.g., mainstream game theory)
 - b. An investment in an agreeable reputation (e.g., game theory again)
 - c. Natural sympathy (e.g., Hume)
2. *Instrumental-cum-moral* Moral action via hypothetical reasoning (e.g., Gauthier)
3. *Non-instrumental*
 - a. Moral action via categorical reasoning (e.g., Kant)
 - b. Social context (e.g., Socrates, Hegel, Marx, Wittgenstein, Habermas)

A radical extension of (1b) and (1c) above [corresponding to (2) in

Table 10.2

] has it that, once an agent recognises the value of cooperation, she has a reason to develop a cooperative *disposition* (as compared to simply acting cooperatively). Gauthier (1986) is the source and argues that the recognition of the value of principled behaviour becomes an independent reason for cooperating. In a sense, it literally pays to be moral. And since the instrumental meaning of 'rational' is grounded on how efficiently higher payoffs are secured, this type of instrumental-cum-moral explanation does not stray far from instrumental rationality. Nevertheless, it challenges rather strongly the conventional instrumental approach by driving a wedge between rational choice and naked preference.

Of course its weakness is, as Hollis (1993) explains, that unless a person undergoes a deeper ontological change (so that she can act on reasons external to her desires)

she will not be able to sustain that *disposition*. Indeed being a person capable of self-restraint (e.g. capable of overcoming the temptation to play R1 when you expect your opponent to play C3) may pay more than being a straightforward payoff maximiser, but the best strategy will always be to dissemble as a principled agent and then cheat. And of course, when people do this, we are back to a world without moral dispositions.

For such an ontological transformation we need non-instrumental reasons for action. One clear suggestion (see (3a) in

[Table 10.2](#)

) comes from Kant (1949, 1959) for whom the distinction between rational and moral choice recedes. Unlike Gauthier, who calls for the development of a moral inclination internal to the agent, Kant's moral psychology distances the agent from her inclinations; instead it empowers her to trump her desires when they are incompatible with a universalisable (moral) principle. In this light, when people cooperate in our game this is seen as worthy action because it is activated by the 'right' motives and independently of consequences; even if players who cooperate gain more dollars. Greater gain is a welcome by-product and not the cause of moral action. However, we are still in the realm of rational action because it is reason which, according to Kant, motivates the will in this particular way.

Finally, we have a melange of explanations (3b) which lead us away from an individualist perspective. Returning for a moment to ancient Greece, Socrates suggests that, prior to action, we ought to ask ourselves: 'How should we live in order to achieve (loosely translated as "good living")?' He suggests that our goal must be a successful life (as opposed to an enjoyable one) and that, crucially, it is through a dialogue with 'others' that we will come to decide whether we have achieved our task. So, while purposefully sidestepping the conceptual minefield of morality and virtue, Socrates introduces 'others' into our calculation of what it is rational for us to do.

More recently, Gilbert (1989) has commented that agents can coordinate their actions (and avoid the temptation to cheat) in interactions like those above, provided they find a mutually beneficial and *shared* line of reasoning. This is different to Hume's natural sympathy argument because it is our reason which is responsible for the coalescence – not our passions. Suddenly, our players see each other as partners, rather than as opponents. Is it insignificant, for instance, that

game theorists always speak of 'opponents' even if a superior, mutually beneficial outcome such as (R3, C3), is available? If players manage to conceive of themselves as one decision-making unit (again, notice the difference with Hume), then cooperative moves in our games cease to be paradoxes in need of de-mystification.

Nevertheless, to sustain this view it is important to explain how it might be rational to conceive of the 'other' as part of a unit to which you belong also. One way to do this is to follow Hurley (1989) in her attempt to establish some Archimedean vantage point from which to judge who can qualify (rationally) as partners. Clearly, expanding the borders of our 'self' to include others is one way in which cooperation in our game can be understood. I have already discussed this extensively in

[Chapter 7](#)

. But, does this type of reasoning not mean that we are expanding the borders within ourselves simply because doing so is an end in itself? Not necessarily.

For the ancient Greeks moral action, as understood by Western philosophy, was not an issue (see Rowe, 1993). In the earlier description of our games I tended to describe the choice of strategy R3 (or C3) in terms of a tension between cooperative (or moral) and self-interested (or instrumental) action. To Socrates this would be nonsense: to choose anything other than R3 when in the R role is shameful – regardless of whether anyone is watching us! The crux in his thinking is the derivation of 'shame' from the realm of the Polis. Indeed. it has nothing to do with morality depending on the extent to

which the self has some natural sympathy for the preferences of others – as Hume (1888) suggested. Socrates might have agreed with Hume that we care about ourselves fundamentally, yet the great difference is that to be rational in Socratic terms is not to be slaves to our preferences but rather to seek our own *eudaimonia*. And, whereas preferences are private, the concept of *eudaimonia* – just like those of ‘truth’ and ‘good’ – is available to all, provided they seek it through dialogue within a community of persons.

So, the reason why we should cooperate is because we will not be leading the good life if we do not. What distinguishes Socrates from Kant, even though their recommendations to our players are identical, is the same point which distinguishes him from Hume: for Kant the perspective from which the right principles are drawn is that of the well-defined individual (who can derive these rules from the data of the game alone, without reference to a social context), whereas for Socrates it is the perspectives of ‘others’ which count. By seeking a clear (i.e., rational) reflection in their eyes – in the eyes of our community—we try to see whether our actions correspond to those which are constitutive of the good life.

Others have followed Socrates down this dialectical path. Hegel’s (1953) conception of a reason which evolves as the ‘self’, rationally reflects on the ‘other’, and ultimately reflects the progress of political society, is but one example. Marx (1963) – a spiritual child of the ancient Greeks and of Hegel – denounced any attempt at defining the meaning of rationality outside the specific social context shaped by the technology and social organisation of the community. Wittgenstein (1953) rejects that action (such as choosing the cooperative strategy in our games) should be informed by exclusive reference to the data of the game, or the mental state of the agents. Instead, he would suggest that the moment

the game is *described*, a process of interpretation of each strategy begins; players try to attach meaning to their available strategies. If they conceptualise strategies R3 and C3 by means of the linguistic signifier ‘cooperative’, then whether they will play them or not depends to a large degree on whether there is an institution of cooperating in the community from which they have been abstracted and which has created their language. Finally, Habermas (1990) completes this Greco-German group with his famous definition of rationality as communicative action.

10.3 The evidence from the laboratory

Let us return to the experiment and see whether it sheds any light on all this. The task of sifting through the six explanations of cooperative behaviour is assisted firstly by the experimental design and secondly by the observed behaviour. Let us start with the former.

Our players were divided into groups (ranging from 8 to 16) and punched their choices simultaneously into a computer terminal without knowing who they were playing against; a computer network assigned them at random to some other player in their group. Moreover, in each round they played against another random draw from the group. The computer did not allow for the same pair to play a game twice in a row. Another constraint to the randomisation of pairs within the four repetitions of the same game was that each player was assigned the role of R twice and the role of C twice. Subjects were made aware of all this at the outset.

At the end of each round players were informed of their opponent’s (or should I say partner’s?) choice, and thus of their score, as well as of the frequency with which different strategies and outcomes eventuated in their group. At the end of the session, their payoffs from each round were summed up and translated into Australian dollars. For instance, if during some round outcome (R1, C3) occurred, then the person with the R role was credited with \$10 while the one with the C role lost \$1. At the outset we guaranteed our subjects a minimum (final) payment of \$10 in order to dispel any fears

that they would make a net loss; nevertheless, this floor never became binding.

Why would players cooperate in this game? As explained in the previous section, the experimental design ruled out explanation (1b) outright: since the games were played anonymously and the chances of meeting the same player in the next round were zero, there is no room for explaining cooperative moves as an investment in reputation. And yet, preliminary experimental sessions (the results of which are not reported here) showed that more than half the strategy choices were cooperative. Why? Since explanation (1b) has been disqualified, perhaps the truth lies within explanation (1a). But then we would be dismissing roughly more than half of our subjects as rationally defective, especially in view of the relatively high stakes involved (recall how a successful cheating move rewarded a subject with \$10). Moreover in every session we ran we found that the total number of cooperative moves was not falling. This result contradicts explanation (1a), according to which experience should teach subjects to

avoid cooperation. Something more subtle and interesting was happening in our laboratory.

The first clue of this emerged when we noticed a quite extraordinary result. As the sessions unfolded, and our subjects gained more experience, a pattern emerged in every session (we ran eight such preliminary sessions featuring the game in

Table 10.1

) comprising three startling features:

- (a) The number of cooperative outcomes (R3, C3) was high but declined steadily;
- (b) Whereas the Rs 'learned' to cooperate less, the Cs cooperated more and more;
- (c) The total number of cooperative choices remained steady (it even showed a tendency to rise).

Note that result (a) is seemingly compatible with the instrumental logic of (1a) and could be taken as evidence that greater experience 'teaches' players to cooperate less (since the correct calculation of strategic advantage should discourage cooperative play). Yet (c) undermines this explanation as the total number of cooperative attempts shows no tendency to fall. The major clue (as well as mystery) lies within (b): Why do the Cs cooperate more as time goes by, while the Rs cooperate decreasingly? And how come the rise in the propensity of the Cs to cooperate is so strong that it cancels out the Rs' tendency to abandon cooperative behaviour, thus producing result (c)?

Could it be that the Rs and the Cs differ in character? Of course not! Recall that the Rs and the Cs are the *same* people who alternate between the two roles constantly. So, how come the same people recognise their strategic position when in the R-role but turn a blind eye to it when in the C-position?

While contemplating this paradox, and before we examine the remaining explanations in

Table 10.2

, it is tantalising to recall the words of the Athenian generals. On hearing the Melian argument as to why it was in Athens' interest that Melos should be treated with respect, they replied: '[W]e know that you or anybody else with the same power as ours would be acting in precisely the same manner' [s104]. Could it be that something similar is happening in the laboratory? Namely, that the Rs have a strategic advantage over the Cs and it is for this reason that the latter exhibit a quasi-moral urge to cooperate (i.e., to choose 3) and to defy the computation of strategic play?

To appreciate the Rs' strategic advantage, observe that they can aim at two birds with one stone. If they choose R1, then they collect 5 if C chooses C1 or, even better, 10 if C tries to cooperate by playing C3. In contradistinction, the Cs cannot do this. Unlike the s, the Cs do not have access to a strategy which *simultaneously* aims at payoffs 5 and 10. Either a C will target payoff 10 (by selecting C1 in the hope that R will choose R3) or C will aim at payoff 5 (by choosing C2). Although the Rs suffer from the

same combination of fear and temptation regarding cooperative play as the Cs (notice that if R expects to choose R3, C has an incentive to abscond by choosing C1), they are in a better strategic position as the Cs. The reason is that failed attempts by the Cs to 'cheat' (i.e., s choosing C1 but the Rs responding with R1) result in a 5 payoff for the Rs and a 0 payoff for the Cs.

Table 10.3

Versions 2 and 3 of the game played by subjects in the laboratory

	C1	C2	C3
R1	5, 0	-1, -1	-1, 10
R2	-1, -1	0, 5	-1, -2
R3	-1, 1	-2, -1	6, 6

Version 2

Instrumental prediction: No R-player will choose R3 although C-players might play C3 if they expect R to choose R3. As the game unfolds, R-players will be playing 3 less often and so C- players, recognising this, will also play C3 less often. The predicted 'attractor' is (R1, C1).

	C1	C2	C3
R1	5, 0	-5, -1	-1, 10
R2	-5, -1	0, 5	-1, -2
R3	-1, 1	-2, -1	6, 6

Version 3

Instrumental prediction: R-players may now play R3 if they expect C3 to be a likely choice of their 'opponent' and fear that C2 is a strong possibility compared to C1. More cooperative play by both Rs and Cs is one possibility as the game evolves. Two other possibilities are (R1, C1) and (R2, C2).

To test the *Athenian hypothesis* – that being in a disadvantageous strategic position makes one prone to a moralistic outlook and, therefore, to quasi-moral deeds which defy strategic calculation – two additional versions of the game in

[Table 10.1](#)

were tried (see

[Table 10.3](#)

).

Version 2 differs only in one small detail from version 1: in the bottom left cell, C's payoff is 1 rather than 10. The meaning of this is that C no longer has a reason to 'cheat'. Whereas in version 1, a C expecting R to cooperate (i.e., to play R3) would be better off selecting C1 (this is what I defined as 'cheating' or defecting), in version 2, C is better off playing C3 (i.e., cooperating) when she or he expects R to play R3. However, the fact that the top-right cell remains the same means that R retains a strategic interest in 'cheating.' In conclusion, version 2 deepens the strategic disadvantage of the Cs. Not only can the R-players aim at their 5 and 10 payoffs simultaneously (which is the case both in *versions 1* and 2) but now Rs have an incentive to cheat whereas Cs do not.

By making the strategic advantage of the Rs over the Cs more pronounced, version 2 offers an interesting test of the hypothesis (inspired by the Athenian generals) that strategic disadvantage makes one more likely to moralise. ...

Lastly, to give the experiment another twist, version 3 added another type of strategic asymmetry between Rs and Cs. By making the Rs lose five dollars every time asymmetrical outcomes (R1, C2) or (R2, C1) occurred, while the Cs continue

to lose only one dollar in those situations, the Rs were given something to worry about; i.e., *some* strategic advantage was returned to the Cs. Instrumental logic suggests that the Rs may, for the first time, have a strategic reason to cooperate since R3 is the only strategy that does not pose the danger of forfeiting five dollars. Given that this is so, the instrumental prediction is of a higher frequency of R3 choices which would also spawn more C3 choices (since the Cs retain the interest, occasioned by version 2, in cooperating more provided they trust that the Rs will do so too).

However, if the hypothesis toyed with here is correct, the total number of cooperative moves may actually fall. This extension of the *Athenian hypothesis* suggests that those who enjoy an improvement in their strategic position will cooperate less, not more. Thus, the Cs, having observed the improvement in their position, may not cooperate more than they did in version 2. Moreover, the improvement in the strategic position of the Cs (and thus the deterioration in that of the Rs) may not be so dramatic as to alter the behavioural pattern as well as the outlook of the Rs. In that case, overall cooperativeness will diminish as a result of the introduction of version 3.

The three versions of the game were tried out in 14 sessions involving 156 volunteers. Most were University students from different faculties of Australian, Austrian, Greek and Hong Kong Universities. A small proportion of participants were professional people, most of whom had University degrees. None had been exposed to game theory before. The sessions were conducted as follows: Players were asked to play four 'warm up' rounds of the 2×2 part of version 1 (i.e. version 1 without strategies R3 and C3) in order to familiarise themselves with simple matrix games. Then they played version 1 four times followed by version 2 another four times. At that point they were made to play the 2×2 part of version 3 (i.e. version 3 without strategies R3 and C3) so as to notice unambiguously the change in the (R1, C2) and (R2, C1) cells of the matrix. Lastly they played the complete version 3 four times. Subjects alternated constantly between the R and the C roles in a manner which ensured that they were in the R role half of the time (the other half they played as Cs). In each round they were asked to punch into the computer their prediction of their opponent's choice and, immediately after, their own choice. Once all choices were registered the computer informed each player of their opponent's choice (and hence of their payoff) as well as the frequency with which each strategy was selected in the group.

Table 10.4

relates the basic results. Concentrating on version 1, the data confirms the preliminary results: Firstly, there is a great deal of cooperative play. Secondly, as anticipated by the *Athenian hypothesis*, the Cs are more cooperative than the Rs (150 as opposed to 134 cooperative moves), something that the Rs had anticipated (the Cs expected cooperative moves by the Rs 177 times as opposed to the Rs who predicted that the Cs would cooperate 190 times). Note also that, in the preliminary experiments (not reported here), in which version 1 was played 16 times (as opposed to the 4 here), the difference in cooperativeness of the Rs and the Cs grew from round to round. The *Athenian hypothesis* (that the strong do what they want and leave the weak to adopt a moral outlook) is reinforced by two

observations from version 1:

- (a) The Rs cheated successfully 70 times (i.e. there were 70 (R1, C3) outcomes) compared to the 56 occasions when the Cs cheated;
- (b) Whereas 70.5 per cent of the Rs who expected cooperation from the Cs (i.e., predicted C3) actually cooperated, the comparable figure for the Cs (i.e., the proportion of Cs who expected R3 and played C3) is a significantly higher 85.9 per cent.

However, the interesting test for this hypothesis comes with version 2 (which enhances the strategic disadvantage of the Rs). Recall that in version 2 a C-player ought to cooperate more the greater her/his expectations concerning the likelihood that

R would cooperate. Remarkably we found that in the four rounds of version 2 (compared to the four rounds of version 1) whereas the Rs reduced their cooperation by 22.4 per cent, the Cs cooperated 46.7 per cent more often! Were they misjudging the Rs? Did they indulge in wishful thinking, expecting the Rs to cooperate more? No, is the answer.

Table 10.4

reports that the Cs actually *foresaw* the reduction in the cooperativeness of the Rs during version 2. Indeed, looking at the data on the number of times Cs expected Rs to cooperate, we find that the Cs expected 22.6 per cent less cooperation by the Rs compared to version 1 (to be precise, from 177 in version 1 to 137 in version 2) and cooperated 46.7 per cent more! Why?

Whatever the reason, the data in

Table 10.4

is striking. Observing the top-right and the bottom-right corners of the matrices corresponding to versions 1 and 2, one notices that, in spite of the significant reduction in cooperative attempts by the Rs, the total incidence of successful cooperation [i.e. the frequency of the (R3, C3) outcome] rises from 75 to 82 while the number of times the Cs were 'cheated' on by the Rs [i.e. the frequency of outcome (R1, C3)] explodes from 70 to 136. Clearly, the propensity of the weaker Cs to move against the calculation of their strategic interest, and oppose the tide of the Rs' increasingly aggressive espousal of R1, contributed to more cooperation (as well as to more 'cheating'). It seems that, even if the Athenian generals were right in saying that the weak moralise because of their strategic weakness, such a perspective may be encouraging the 'weak' to act in a manner that is conducive to more cooperation in the face of more aggression by the 'strong'.

Turning to version 3, the evidence from the laboratory shows that, once a pattern favouring the Rs has been established, it is difficult to dislodge. The pattern in question of course is the *increasing* tendency of the Rs to play R1 (aiming for the non-cooperative \$5 and \$10 payoffs simultaneously) at the expense of the Cs who continue to opt for C3 thus collecting negative payoffs more often (because of the s' convergence onto R1).

Table 10.5

confirms that in version 3 the occasions of 'cheating,' outcome (R1, C3), increased further (from 136 to 146 instances); just as they rose when version 1 gave way to version 2 earlier. However, this time the further shift against the Cs was not accompanied by more cooperativeness by the Cs.

Table 10.4

Aggregate behaviour: the stubbornness of cooperation

'learning' their lesson, as the instrumental view would have it, why did the Cs cooperate 214 times when they only expected cooperation 146 times (instrumental logic suggests that, in versions 2 and 3, players in the C-role will cooperate *only* when they anticipate cooperation)? Could it be that, despite the slight improvement in their strategic position, they remained in a decidedly weak position and therefore continued doing what the weak do (that is, act quasi-morally)? Perhaps the only effect of the slight shift in strategic strength towards the Cs was that, in version 3, the Cs expected the Rs to cooperate 6.6 per cent more often (compared to version 2) and responded to that by cooperating a mere 2.7 per cent less often.

So far our experiment has disqualified two of the explanations of quasi-moral acts listed in

Table 10.2

and added the *Athenian hypothesis* to it. Explanation (1b) has been ruled out by the experimental design and (1a) receives very little empirical support given that experience does not discourage cooperative moves by the Cs as it would if such moves were mere execution/calculation errors. This leaves explanations (1c), (2), (3a) and (3b) in play. If we are to discriminate between them as well as to contemplate the *Athenian hypothesis*, we need to look at the data more closely.

Table 10.5

attempts to delve more deeply. The first column copies the raw number of observed cooperative attempts from

Table 10.4

. The second column (labelled P3P3, standing for 'predicted strategy 3, played strategy 3') tells us how many times a cooperative strategy was played *when the person who played it was anticipating cooperation*. The third column reports on the frequency of 'sacrificial cooperation' (i.e., occasions when a player cooperated *even though she/he did not anticipate cooperation*). The column labelled 'continuing cooperation' counts the occasions in which a player, having experienced mutual cooperation [i.e. outcome (R3, C3)] in the previous round cooperated again in the next round. Finally the 'cheat' column relates the number of cases in which a player expected cooperation (that is, expected the third strategy) but chose not to reciprocate and played strategy R1 or C1 instead – the 'cheating' option. Lastly, note that no such data is applicable to versions 2 and 3 because in those versions C-players predicting R3 are best off replying with C3; that is, they have no incentive to cheat.

Before looking at the data in

Table 10.5

, let me rehearse the main puzzle once more: Why do the Cs cooperate more than the Rs in version 1? And why is it that, in spite of the clear move of Rs away from their cooperative third strategy

R3, the Cs espouse C3 with such fervour *even though they anticipated correctly that the Rs' would play R3 less often*? One thing is clear: there can be no instrumental explanation of this observation along the lines of explanations 1(a) or 1(b) in

Table 10.2

. To put it differently, no conventional game theoretical explanation is forthcoming. Given that the Rs are playing R1 more and R3 less often, the Cs ought to be playing C3 less and less. And yet they do the opposite. Perhaps the clues lie in

Table 10.5

Table 10.5(a)

tells us that in version 1 of those Rs who predicted that their 'opponent' would cooperate, 47.36 per cent of the time they 'cheated'. Compare this to the figure for the Cs: only 37.85 per cent of the time did they 'cheat'. Moreover 52.63 per cent of the Rs

who predicted cooperation cooperated, compared with 62.15 per cent for the Cs. As we have seen in the context of

[Table 10.4](#)

, this significant difference cannot be explained straightforwardly by any of the explanations in

[Table 10.2](#)

. Once more the Athenian general's retort to the Melian argument – namely that morality is for the strategically weak – proves inviting.

[Table 10.5a](#)

The row players' actions and beliefs

<i>Row players</i>	<i>Row players</i>				
	<i>Cooperated – i.e., played R3</i>	<i>P3P3–i.e., predicted C3 and then played R3 (in brackets the number of times Rs expected C3)</i>	<i>Sacrifice – i.e., cooperated when they did not expect C3</i>	<i>Continuing cooperation – i.e., play C3 immediately after an (R3, C3) outcome</i>	<i>Cheat – i.e., predicted C3 and then played R1</i>
Version 1	134	100 (190)	34	33	90
Version 2	104	87 (231)	17	21	144
Version 3	84	67 (231)	17	20	164

[Table 10.5b](#)

The column players' actions and beliefs

<i>Column players</i>	<i>Column players</i>				
	<i>Cooperated – i.e., played C3</i>	<i>P3P3–i.e., predicted R3 and then played C3 (in brackets the number of times Cs expected R3)</i>	<i>Sacrifice – i.e., cooperated when they did not expect R3</i>	<i>Continuing cooperation – i.e., play C3 immediately after an (R3, C3) outcome</i>	<i>Cheat – i.e., predicted R3 and then played C1</i>
Version 1	150	110 (177)	42	39	67
Version 2	220	117 (137)	97	43	N/A
Version 3	214	130 (146)	81	36	N/A

But could there be a more conventional explanation of what is going on in our laboratory? In version 2, 85.4 per cent of the Cs who expected cooperation cooperated. By version 3, this figure had become 89 per cent (while only 29 per cent of the Rs who expected cooperation cooperated.) This makes perfect instrumental sense, as it is instrumentally rational for the Cs to play C3 *if they anticipate* R3. Still, it does not explain why, in version 2, the frequency of cooperative moves by the Cs increased even though they expected less cooperation rather than more. And it does not explain why the total number of cooperative moves (of Rs and Cs taken together) rose significantly from version 1 to version 2 (and remained almost constant in version 3) on the back of the incredible cooperativeness of the Cs who acted in this manner even though they anticipated and experienced very little (relatively speaking) cooperation from the Rs. Could the answer be found in explanation (2) of

[Table 10.2](#)

?

For Gauthier (1986), rational agents should foresee that unconstrained maximisation

will, at best, lead them to the rather poor payoffs resulting from strategy combinations (R1, C1) or (R2, C2). By contrast, if they could acquire a cooperative *disposition* they would be captivated by strategies (R3, C3) and, as a result, boost their rewards. The problem here is that Gauthier's logic should, if correct, prevail regardless of whether a player occupies the R or the C roles. Since they know well that they are alternating between the R and C positions, the strategic asymmetry between the Rs and the Cs should make no difference: they should act as constrained maximisers who, regardless of role, opt for the third strategy.

Another explanation which seems unsatisfactory is (3a) of

Table 10.2

. If our subjects' cooperative moves are due to some type of categorical reasoning a la Kant, why did they act aggressively when in the R-role and more cooperatively when in the C-role? Surely a universalisable imperative ought to be just that; an imperative demanding the agents' principle trumps calculations of strategic advantage consistently. Indeed, our subjects showed no tendency towards categorical principles; as we have already seen, when in the R-role the proportion of those who cooperated when anticipating cooperation fell from 52.6 per cent in version 1 to 37.66 per cent in version 2. Hardly the type of behaviour that Kant would condone.

We are left with explanations (1c) and (3b). According to the former, the role-specificity of cooperative behaviour must be due to the different passions (e.g. one for money, another for equity etc.) brought into the laboratory by the players. What distinguishes this Humean interpretation from the conventional rational choice model is that Hume makes no assumption regarding the commensurability of these passions. It may very well be the case that the agent is torn in a manner which cannot be easily settled by means of a clean ordering of preferences. This might explain why sometimes a player cooperates while at others she/he does not, even when the objective (i.e. the strategic) data remain unchanged. And (if Hume was right) this has nothing to do with the person's rationality; reason (the passions' slave) should not be blamed for unruly passions.

A similarly contextual interpretation is offered by the melange of thinkers under (3b). The crucial difference is that they, unlike Hume, do not conceptualise

motivation as clinically separable between an impartial, static, asocial reason on the one side and the non-rational passions (in which the will lies) on the other. For Socrates, Hegel, Marx, Wittgenstein and others, people create their reason as they create the rest of their lives: socially. For them, the three strategies in our games are first *interpreted* in terms of social data they have brought with them, and then selected on the basis of that interpretation. And since interpretation is inherently haphazard, due to persons' diverse social locations, different things and ideas motivate different people in our laboratory – regardless of our attempts to impose on them (through the experimental design) uniform ends. The different propensities of players to cooperate depending on whether they are assigned the R or the C role is thus explained by the interpretative differences in motivations (their 'ideology' as Marx would insist) engendered by the payoff matrices' asymmetries. Indeed, subjects may plausibly come to the conclusion that, when a C-player, it is *better* to be more cooperative than when an R-player; where *better* is of course to be understood independently of instrumental concerns.

The most remarkable observation has been left last: as time went by and our players moved from one version to the next, the frequency of sacrificial acts of cooperation (that is, someone cooperating even though she/he does *not* expect the other player to do likewise) follows different patterns depending on whether the player is in one or the other role. The main point here is that *it never makes sense to cooperate against an uncooperative opponent*; in fact, it makes no sense either from an instrumental

perspective or, indeed, from a non-instrumental one. For example, neither Gauthier's hypothetical nor Kant's categorical imperative in support of playing R3 and C3 recommend to players that they should cooperate no matter what. Hume too, I dare presume, would not sanction the concept of natural sympathy towards those who take advantage of one.

Table 10.5

contains some fascinating data on sacrificial cooperation. When in the R-role, such acts (numbering 34 over the four rounds) constituted 25.4 per cent of all cooperative moves (i.e., R3 choices). In version 2 their number dropped to 17 (16.3 per cent of all R3s) and in version 3 such acts fell further to 10 (11.9 per cent of all R3s in version 3). Compare this to what *these same people* did when in the C-role. In version 1 sacrificial cooperation was adopted 42 times by the Cs (28 per cent of all C3s). In version 2 that raw figure more than doubled to 97 occurrences (and to 44 per cent of all C3s). Only in version 3 was the surge checked (the raw frequency fell slightly to 81, i.e., 37.9 per cent of all C3s). Such data strengthens further the suspicion which motivates this chapter's argument; namely that the Athenian generals were not totally wrong when they said that a weak strategic position causes agents to adopt quasi-moral behaviour.

According to their account, the Cs plunged into a sea of sacrifice when their strategic position became even worse than it was at the start (i.e. when version 2 replaced version 1) and only held back slightly when their position was strengthened somewhat (i.e. with the introduction of version 3). Can this phenomenon be explained in a manner compatible with Hume? Doing so would necessitate his theory of conventions. When the passions under-determine choice, agents achieve consistent behavioural patterns through trial and error. Thus,

a social convention emerges which agents learn to observe because doing so reduces wasteful indeterminateness. So, although it is within reason to follow conventions, the specific convention people end up following is not uniquely rational. As Hume (1888) put it: 'Tis not, therefore, reason, which is the guide of life but custom.'

In our laboratory our subjects generated endogenously a social convention according to which those in the better strategic position (the Rs) make full use of their strategic possibilities whereas the less fortunate make more sacrifices, under-utilise their strategic position, and rise onto the higher moral ground. Thus, it is possible to imbue Hume with the Athenian hypothesis that strategic weakness evokes a passion for making sacrifices for the common good.

Naturally, Hume's separation of reason from society is highly controversial and not everyone's favourite move (see Varoufakis, 1991, and

Chapter 4

of this book). Under (3b) we find a host of alternative explanations of the dynamics of cooperative behaviour as socially and historically determined. Socrates introduced the idea of reason as argument in the public sphere and Hegel saw reason develop within the evolution of social norms (unlike Hume's reason which is a disinterested, unchanging part of the agent's mind). Marx tied that evolution to the history of the social organisation of production and Wittgenstein introduced us to the mutual constitution of action and structure in the practices of a community of persons. Any one of these traditions could shed light on the social convention which was generated endogenously in our laboratory: namely that, *those in the strategically weaker position develop a tendency towards expectations, actions and rhetoric which can be seen as morally motivated*.

However, what is perhaps a more practical lesson is that the usual categories of moral motivation (of which

Table 10.2

is a scant example) should include a brief mention of an ancient dialogue between

the generals of an invading army and their desperate, yet philosophically adept, victims.

10.4 Back to the Melians: imperialism and the moral authority of the weak

The actual events depicted in the *Peloponnesian War* make the questions in this chapter look rather academic. The Melians' plea was doomed from the start, regardless of its elegance, rationality or moral content. It failed because Athens did not aim at a reputation for magnanimity in victory. Indeed, its objective was the opposite: a reputation for ruthlessness towards those 'allies' who absconded its sphere of influence; an ironic twist on explanation (1b) of

[Table 10.2](#)

The Athenians were disarmingly honest on this. Asked why they could not accept an independent, yet friendly, Melos they replied: 'No, because we are not injured by your hostility; rather we are worried that, if we were on friendly terms with you, those whom we have already subjugated would regard this as a sign of weakness in us, whereas your hostility is evidence of our power' (Thucydides, *History of the Peloponnesian War*, Book 5, 95). Melos' fate is testimony to the impotence of abstract morality against the logic of imperialism. Yet it

offers no evidence that moral rhetoric is irrelevant: at least one Athenian (that is, Thucydides) was impressed by their argument.

Our experiment has contributed three thoughts to the assessment of that argument: (a) Deeds with a moral appearance are irreducible to sophisticated expedience: In a laboratory setting which abolished the incentive to act cooperatively because of any anticipated gains to be had from a reputation for (or appearance of) virtue, the will to cooperate proved remarkably resilient (see

[Tables 10.4](#)

and

[10.5](#)

).

(b) A moral disposition is unlikely to be acquired instrumentally, as Gauthier (1986) would have it. Additionally, and contrary to Kant's view, principled behaviour seems neither universalisable nor independent of strategic motivation.

(c) Those with the weaker strategic role were expected to indulge more often in good deeds (cooperation in our case) *regardless of any calculation of strategic advantage*. Indeed such expectations were confirmed in practice as the group's behavioural pattern evolved.

Thus, the experiment does not lend straightforward insights to the Melians' case. On their side they have the first finding: there is room for a moral stance such as theirs irreducible to strategic pursuits of self-interest. However, the next two are less sympathetic. Finding (b) reinforces scepticism about the chances that Athens' imperialist plans would be shelved, trumped by moral (internal or external) reasons which were supposed to have been activated by the Melian representative's fiery speech. To make things worse, had the Athenians had access to finding (c), they would feel vindicated for having dismissed the Melian speech as the inevitable moralising of the feeble.

In many ways, the Athenian general foreshadowed this by implying that a rerun of history following a reversal of fortunes would offer conclusive evidence on the insincerity of the Melian position: '[W]e know that you or anybody else with the same power as ours would be acting in precisely the same manner' [s104]. Was he right to think so? Unlike history, which is not obliging on this, the verdict from the laboratory leans in his favour. Many (and on occasion most) of our participants swapped happily their

cooperative choices for strategic aggression when they moved from the weaker to the stronger role.

There is one perspective which has so far only been foreshadowed conspicuously by a reference to Aristotle in the Introduction. It takes its strongest form in the words of Nietzsche:

there is master morality and slave morality ... those qualities which serve to make easier the existence of the suffering will be brought into prominence and flooded with light. Slave morality is the morality of utility.

(Nietzsche, 1973)

As the Melians tragically found out, and our experiment confirmed, a world systematically segregated between the dominant and the lesser social roles may indeed evolve into a world of slave and master moralities. This is where Aristotle and Nietzsche were right. Where I hope they were wrong is in their conviction that such segregation is due to *natural* differences between people. Thankfully our experiment casts doubt on this interpretation:

Table 10.5

(especially the column on 'sacrificial cooperation') shows clearly that one's moral disposition depends on one's *social location* rather than on an intrinsic strength or weakness (since the Rs and the Cs were the same persons). Therefore, Nietzsche's separation between the weak and the strong may well be as artificial in society as it was in our laboratory; in which case all that is needed to undo it is a re-designed social context. If this is so, there is hope that Nietzsche was also wrong to think that the Will to Exploit is 'a consequence of the Will to Power, which is after all the Will to Life' (Nietzsche, 1973). The resonance of the Melians' speech through the centuries augments that hope.

10.5 Epilogue

When I was designing the experiment mentioned above, some time in 1995, I had already lost hope that neoclassical economists could be shaken out of their set, complacent ways by the sheer force of empirical evidence that violated their prejudices. The reason I went ahead with it was pure curiosity on my part to see if game theory could be used, in an experimental framework, in order to show that social roles and conventions can evolve, in a short space of time, in such a way as to engender systematic patterns of discrimination. The whole idea behind the experimental design was to ensure that any observed discrimination between groups could not be 'blamed' on the character of these groups' members. To achieve this, I ensured that the very same people made up both groups (*R*-players and *C*-players), alternating between the two roles in different rounds.

That one of these 'groups', or roles to be more precise, should discriminate against the other, while the latter fell back on 'moralistic' behaviour that resonates with the Melians' rhetoric, was quite remarkable. At least, so thought the editor of a leading analytical philosophy journal, *Erkenntnis* – the journal founded by the Vienna Circle many decades ago – which published the resulting paper. When *Erkenntnis* published it, in 1997, I did not exactly expect economists to take much notice; after all, they could not be reasonably expected to read a philosophy journal, especially when the paper was written in a language geared towards philosophers.

Nevertheless, the power of the experimental finding above seemed to me substantial enough to attempt, for one last time, to impress it upon my neoclassical colleagues. To do so, I re-designed the experiment and ran it again, this time in a manner that conformed fully to the specifications that a neoclassical economist would expect of both the experiment and the manner in which it was written

up. The result was a sequel that was published in 2002 by the *Economic Journal* (jointly authored with Shaun Hargreaves-Heap). The contents of that paper are re-

produced, and re-appraised, in the following chapter.

Note

¹

All translations from the Greek text are the author's.

11 Evolving domination in the laboratory

The spontaneous creation of hierarchies and the patterned beliefs that support them

11.1 Prologue

11.1.1 Background briefing

The last paragraph of the previous chapter explains fully the motivation behind the experiment to which the present chapter is dedicated. Taking neoclassical economists on their word that they are staunch empiricists, some time in the 1990s I decided to design experiments which raised empirical doubts about the solidity of the foundations underpinning neoclassical dogma.

In particular, I set out to design an experiment that illustrated the remarkable capacity of humans to create, almost from nothing, patterned behaviour that discriminates on the basis of arbitrary characteristics which, according to neoclassical theory, should make not an iota of a difference. Put differently, I endeavoured to show to my neoclassical colleagues that it is perfectly possible to have systematic behavioural patterns which constitute a vicious form of discrimination that does not, nonetheless, reflect anything 'real'.

Aware of the conservative turn amongst 1990s neoliberals, which caused them to think of sustained race and gender discrimination as some sort of 'proof' that women and blacks were somehow 'challenged' (remember the awful book entitled *The Bell Curve?*), I thought it would be interesting to see if sustained discrimination could evolve in a laboratory between groups that were virtually identical. The fact that such a result would never square with the neoclassical model of men and women made it an exciting proposition.

The previous chapter presented the first such experiment, as published in a philosophy journal. As I explained in the last chapter's epilogue, I then set out to design an even more powerful experiment that I intended to publish in a leading economic journal. This would also constitute an experiment: not an experiment in how lay people behaved in an experimental laboratory but, rather, an experiment to test how genuine neoclassical economists are when they claim to be empiricists who would never deny the facts' primacy over their theoretical prejudices.

11.1.2 The rest of this chapter

1

Many economic interactions mix mutual benefit with a measure of conflict. For instance, when two people trade, there is often more than one price where both will benefit. The high end of the range favours the seller while the lower advantages the buyer. So, when they settle on a price and trade, they unlock a mutual benefit and resolve a potential conflict. The hawk–dove game (HD) captures these elements, albeit in a rather simple way as each player only has a choice between being a hard bargainer (a hawk) and a soft one (a dove). Nevertheless, this is why it is regarded as one of the classic games of social life and why it is important to be able to predict behaviour in this game.

Prediction, however, is difficult in the HD game for reasons that relate to some fundamental issues in game theory. The game has multiple Nash equilibria and the equilibrium selection problem is not readily solved, if we stick with the mathematical description of the game, by an appeal to salience. The symmetrical solution, for instance, echoes the symmetry of the game, but it is not a Nash equilibrium and so does not seem a good candidate for salience. Likewise, the two pure strategy equilibria are symmetrical with one another and so the appeal of one looks as strong as the other.

It is possible, nevertheless, that a factor that is extraneous to the mathematical

description of the game might make one of these asymmetric equilibria salient. Indeed, evolutionary theorists argue that extraneous factors which distinguish between the players and which are common knowledge can 'seed' conventions which advantage one type of player relative to another (e.g., Sugden, 1986; Weibull, 1995; see Lewis, 1969, on conventions).

Others find this explanation of equilibrium selection implausible because the inequalities in outcome are supported only by convention and owe nothing to power, ability, or principles of fairness, etc. While some evolutionary theorists concede that principles of fairness may play a role in equilibrium selection in such games, they also sometimes follow Hume (1888) and argue that these ideas of fairness themselves develop out of the emerging conventions. Thus a convention that evolves in the play of HD may come to be associated with a set of self-validating normative expectations regarding what is fair. These ideas may then come to affect behaviour in other games (see Sugden, 1986, 2000). The main purpose of this chapter is to see whether these processes of convention and idea formation occur in simple experimental games.

First, the experiment tested for whether a convention emerges in the HD game when players are given a piece of distinguishing extraneous information. In particular, players were given either a red or blue identifying colour in the experiment before playing HD and I tested whether the subsequent behaviour was consistent with people following a convention founded on this initial arbitrary colour assignment. Of course, the distinguishing features that might be used in social life are liable to be more complex in origin than this. Nevertheless, it is helpful to know whether conventions can arise in this rather simple experimental setting as it gives an insight into whether the same kind of mechanisms could underpin the generation of conventions in society more generally.

Second, the chapter investigates how ideas of fairness associated with the evolutionary emergence of a convention in one game might affect play in another game.

There are two possibilities here. The principle of fairness generated in one game can act as an equilibrium selection device in other games. Alternatively, these ideas of fairness could feed into a new equilibrium concept: that is, the players' concern to be fair may support non-Nash equilibria in these games. For example, it is sometimes argued in behavioural economics that the selection of the cooperative strategy in the prisoners' dilemma game can be explained through the introduction of 'psychological' payoffs (see

Chapter 8

, Rabin, 1993; or Sugden, 2000 for a similar idea). These are payoffs that are distinguished from the material ones captured in the standard game theoretic representation of an interaction. They arise because players hold beliefs about the fairness of any material outcome which affect their assessment of it.

If fairness does motivate in this way, then it becomes important to understand how people come to have ideas regarding what fairness is: What does it consist of? And when does it apply? The latter is important because these theories also typically generate multiple equilibria and so pose the same question regarding equilibrium selection. To throw light on this problem too, the chapter also reports on an experiment which begins to address such questions.

In particular, the HD game is amended by adding a third cooperative strategy.

²

The amended game is labelled as the hawk–dove–cooperate game (HDC). If both players select the 'cooperative' strategy in this new game, the outcome is symmetrical and Pareto-dominates all three of the game's Nash equilibria. However, the cooperative strategy is not part of any equilibrium according to either standard or evolutionary game theory and, from these perspectives, the new game is strategically the same as HD. Mutual cooperation is, however, a 'fairness equilibrium' in the sense that both Rabin (1993) and Sugden (2000) suggest. The concern here is whether (i) the mutual

cooperation outcome persists in repeated play and (ii) whether this fairness equilibrium was still selected if, *outside the HDC game*, players have experienced a convention which gives one of them an arbitrary advantage. The thought here is that, if Hume's ideas are right, then the ideas of fairness associated with an asymmetric convention in the play of HD will militate against the symmetric fairness equilibrium of mutual cooperation when HDC is played.

Thus, the chapter makes two contributions to the conventional neoclassical literature: It reports on an experiment that is designed to test (a) for the emergence of a convention based on arbitrary colour assignments which enables equilibrium selection in the HD game, and (b) for the endogenous generation of normative expectations in HD which affect play in HDC. The former addresses a prediction in evolutionary game theory; the latter addresses some particular concerns in behavioural economics with respect to the formation and influence of 'psychological' payoffs. The organisation of the paper is as follows.

Section 11.2

sets out and considers the two games in more detail.

Section 11.3

describes the experiment.

Section 11.4

gives the results,

Section 11.5

offers an interpretation and

Section 11.6

concludes.

11.2 The hawk–dove game and an amended version

Table 11.1

presents an HD game. There are two common analyses of this game when it is played repeatedly and anonymously: the standard (or conventional) approach and an evolutionary version.

Standard game theory assumes fully rational agents and finds that HD has three Nash equilibria, two in pure strategies (h, d) and (d, h) and one in mixed strategies ($p = 1/3$, where p is the probability of an h choice).

The evolutionary approach, on the other hand, assumes non-rational players who gravitate toward the strategy with the highest payoffs. In the biological interpretation of an evolutionary process, the gravitation occurs because high payoffs confer reproductive success; whereas in the social interpretation of the process, it happens because people learn from the success of others. It is helpful to consider two possible types of evolution:

One-dimensional evolution: This applies to an homogeneous population. Since
(all members are identical in every way, the evolution of strategies is the same for
i all members.
)

Two-dimensional evolution: All members are identical, with one small
ii exception. Some have one arbitrary feature, the remainder the other. This
) difference, though arbitrary, endows the evolutionary process with a second
dimension because the fact that each player possesses one of two distinguishing
(and observable) features makes it possible for individual behaviour to be
conditioned on one's own feature (as well as on the feature of one's opponent).
The result is that the strategy which gathers popularity among members of one
group may be different from that which is established in the other.

Under one-dimensional evolution, there is a unique evolutionary equilibrium: *the proportion or probability (p) of players choosing h equals $1/3$* . This follows because the average return to a person playing h will be greater (less) than playing d for any value of $p < 1/3$ ($p > 1/3$). Consequently, more (less) players will opt for h if $p < 1/3$ ($p > 1/3$) and p will rise (fall). Therefore, p will only be stable when it equals $1/3$, a value which coincides with the Nash equilibrium in mixed strategies.

With two-dimensional evolution, there are two evolutionary equilibria. Suppose the population is divided into two equally-sized groups by an arbitrary colour identification: members are somehow labelled either blue or red. In meetings between players of different colour the two evolutionary equilibria are: '*red plays h and blue plays d* ' or '*red plays d and blue plays h* ' (see Weibull, 1995, and Friedman, 1996).

The key to this result is that strategies can be conditioned on colour in cross-colour meetings. Suppose that, at the outset and for no particular reason, the frequency of h -play by blue people falls below $1/3$ (and happens to be less than the frequency of h -play by the reds). Then red persons will discover that, when matched against a blue person, the return to h exceeds that of d and thus h -play among red people will increase. This will reinforce the relative attractiveness of d -play for blue people in cross colour encounters. In the end, all blue players will be playing d and all red players h .

Meanwhile the unique evolutionary equilibrium for meetings between players of the same colour coincides with the one-dimensional equilibrium ($p = 1/3$).

Table 11.1

Hawk–dove game

	h	d
h	-2, -2	2, 0
d	0, 2	1, 1

The evolutionary equilibria in mixed colour meetings that result in (h, d) or (d, h) can be interpreted as conventions (see Lewis, 1969). Indeed, they constitute a form of discriminatory convention in the sense that they assign each person, on the basis of his or her colour, to either the hawkish or dove-like role

and this results in people of one colour enjoying much higher payoffs than those of the other, for reasons which have nothing to do with superior rationality, information or contribution.

One objective of the experiment is to test for whether a discriminatory convention of this sort develops when each player is identified by an arbitrary blue or red colour. We call this the *discrimination hypothesis*. The null hypothesis, supported by standard game theory and one-dimensional evolution, is that colour labels will *not* influence behaviour. The alternative hypothesis, supported by two-dimensional evolution, is that players will, eventually, make use of the extraneous information of colour labels to build a discriminatory convention.

The second game (HDC) in the experiment is set out in

Table 11.2

. The original HD game has been amended by the addition of a third 'cooperative strategy', c , for each player. This third strategy is not part of any equilibrium: it will not be played in a repeated setting according to standard game theory and will disappear in the evolutionary version.

Nevertheless, there is some experimental evidence (see Camerer and Thaler, 1995,

for a survey) suggesting that strategies similar to *c* survive (e.g. the cooperative strategy in the prisoner's dilemma).

Table 11.2

The hawk–dove–cooperate game

	<i>h</i>	<i>d</i>	<i>c</i>
<i>h</i>	−2, −2	2, 0	4, −1
<i>d</i>	0, 2	1, 1	0, 0
<i>c</i>	−1, 4	0, 0	3, 3

One explanation of the persistence of cooperative play in interactions like HDC turns on the identification of ‘psychological’ payoffs that come from the symbolic properties of an outcome (its ‘fairness’, ‘goodness’, etc). For example,

Rabin (1993), whose model we studied extensively in

[Chapter 8](#)

, assumes agents who derive utility not only from expected monetary returns but also from a perception that they acted fairly. In his account, the perception of fairness (and hence the psychological payoff) depends on reciprocating ‘kindness’ (or ‘unkindness’). In order to make such judgements, each player needs to form second-order beliefs regarding what his or her opponent expects him or her to play. So, for instance, suppose Cressida is playing HDC against Troilus and contemplates playing *c* because she predicts Troilus will also play *c*. Her utility payoff from outcome (*c*, *c*) varies depending on what she thinks about Troilus’s *motivation* for playing *c*. ‘Is Troilus about to play *c* by accident? Or is he also expecting me to play *c*?’ In the latter case, Troilus’s choice of *c* contains a measure of kindness to Cressida: given his second-order beliefs that Cressida was going to play *c*, he *could* have collected payoff 4 (by playing *h*) but settled for payoff 3 and this enables Cressida to enjoy 3 rather than −1. In analogous manner when she plays *c*, expecting Troilus to play *c*, she also shows kindness to Troilus. When kindness is reciprocated in this way, Rabin argues that Troilus and Cressida both enjoy a ‘psychological’ payoff and when these payoffs are suitably weighted with the material ones, it is possible for (*c*, *c*) to become what is called a ‘fairness’ equilibrium.

8

If the reader wants a refresher course on fairness equilibria, a re-read of

[Chapter 8](#)

is recommended. The point to note here is that Rabin’s theory depends on his definition of ‘kindness’ shown by Troilus to Cressida and vice versa. Rabin assumes that Troilus’s perceived kindness depends on a comparison of Cressida’s actual payoffs from a strategy relative to some assumed reference point. This reference point is given exogenously and defines, in effect, an entitlement for Cressida. When Troilus enables Cressida to obtain something more than this entitlement, he is being ‘kind’. My suspicion is that when people are motivated by such ‘psychological’ payoffs, perceptions of entitlement may be formed in a more complex manner than this; and this is why we have included this game in the experiment.

In particular, it seems of interest to pursue an argument from Sugden (1986) which suggests that ideas regarding what is ‘fair’ or ‘just’ may evolve endogenously in the course of social interaction. Sugden follows Hume (1888) by suggesting that, when a convention emerges in a game like HD, it can induce a set of supporting normative ideas: that is, ideas that make the arrangement seem ‘just’ or ‘fair’ or some such. It is as if people find it difficult to accept that the convention is in some sense arbitrary while also being discriminatory. ‘*Red plays hawk and blue plays dove*’ would perform just as well as a convention as ‘*blue plays hawk and red plays dove*.’ But the selection of one

of these conventions makes a big difference to who receives the most benefit and this seems to cause dissonance. So people remove the dissonance by finding, or inventing, additional principles that will justify the actual convention because it is 'just', 'fair' or some such. If this is the case, then it seems that play of the HD game may induce different ideas regarding entitlements to the play of the HDC game. This is because a convention in HD is inherently discriminatory while it seems from earlier experiments

that people are attracted (possibly on grounds of fairness) to the symmetric (c, c) outcome in games like HDC.

Such a tension between discriminatory and symmetric ideas regarding what is 'fair' or 'just' could make the play of these games sensitive to the order in which they are played. For example, when HDC is played first, a discriminatory convention is less likely to emerge than when HD is played first. This is because the symmetric ideas which may be encouraged by the presence of the cooperative strategies in HDC could inhibit the growth of the discriminatory convention in the play of HD. Likewise, the symmetric (c, c) outcome is less likely to occur in HDC when it has been preceded by HD as compared with experiments in which subjects played HDC first. This is because the discriminatory ideas that might be encouraged in the play of HD could carry over to the play of HDC and inhibit symmetric cooperation. This is the second hypothesis of this chapter which I refer to as the *sequence hypothesis*.

To be specific, the null hypothesis here is that the sequence of play of HD and HDC makes no difference to behaviour in either game. The alternative hypothesis is that a discriminatory convention is more likely to emerge when HD is played first and that mutual cooperation will be different when HDC is played second. This is supported by the idea that people are motivated by 'psychological' payoffs and that the perceptions of entitlements which influence these payoffs depend both on the presence of extraneous information and can be generated endogenously. The comparison with standard game theory and Rabin (1993) is instructive. Since neither standard game theory nor Rabin's theory has a theory of equilibrium selection to offer us, neither makes a prediction regarding an order effect. So if there is an order effect, then neither standard game theory nor Rabin (1993) can explain it.

11.3 The experiment

Four treatments were used to test the two hypotheses. The subjects played each of the two games (HD and HDC) 32 times under quasi-random matching in all four treatments. The treatments differed in two ways: in terms of (a) whether or not players were labelled as blue/red, and (b) whether the 32 rounds of HD preceded, or followed, the 32 rounds of HDC.

In eight sessions no information about individual opponents was provided. We shall refer to them as the No-Colour treatment. In another 24 sessions, the Colour treatment, players were assigned a colour label at the beginning of the session and were informed of the colour label (blue or red) of their opponent. It is by observing behavioural differences between the Colour and No-Colour treatments that we test the *discrimination hypothesis*.

In 16 of the 24 Colour sessions the 32 rounds of HD preceded the 32 rounds of HDC (the **HD-HDC-Clr** treatment). In the remaining eight the order of play was reversed (the **HDC-HD-Clr** treatment). Similarly in four of the eight No-Colour sessions HD preceded HDC (the **HD-HDC-NClr** treatment) while in the remaining four No-Colour sessions HDC was played first (the **HDC-HD-NClr** treatment).

Appendix 11.1

offers full details. It is by observing differences in the pattern of play between the **HD-HDC-Clr** and the **HDC-HD-Clr** treatments that we test the *sequence hypothesis*.

11.3.1 The experimental design

The 640 subjects came mostly from the student population at the University of Sydney over a period of two years. The group size in each of the sessions varied from 16 to 26 (see

[Appendix 11.1](#)

for details). Once seated in front of their terminal, they were asked to consult on-screen instructions and to ask questions of clarification.

The instructions informed players of the following: the total number of rounds (64); the payoff matrix of the first game (either HD or HDC); that the game would be amended after 32 rounds to another game (without telling them what the emendation would be) which would also be played 32 times; that at the end of the session each player would collect in Australian dollars the sum of her or his numerical payoffs from each round;

that one player would win an additional A\$10 from a lottery at the end of the session in which his/her chances would be proportional to how many correct predictions of his/her opponents' choice he/she made; that in each round they would be drawn at random against any player in the group (regardless of colour in the 'Colour' treatments) *except that they would never be drawn against the same player twice in a row.*

10

Following a dry run of four rounds of the first game,

11

the session proper commenced. In the Colour treatments, the colour labels were distributed just before the dry run took place. (Note that the on-screen instructions made no mention of colour labels.) An instructor in full view of players showed them a pack of cards equal in number to that of players. One side of each card was white and the other was either blue or red (half of the cards were blue and half were red). To guarantee that the randomness of the colour distribution was common knowledge, the pack of cards was shuffled in public view. Then the instructor walked over to each subject inviting him or her to pick one at random (before choosing a card subjects could only see the white side of the cards on offer). Once they had collected their coloured card, their screen requested that they punch in 'b' if their card was blue and 'r' otherwise.

Since the games were symmetric, and in order to avoid introducing a second discriminant (namely, a row or column) which could have given rise to four-dimensional evolution, in all treatments players were told that they were choosing among the rows.

12

In each round subjects had to make two decisions. The first was to predict the strategy which their opponent would select in that round. The purpose of this was to gauge the first-order (predictive) beliefs of subjects for later use (see

[Section 11.5](#)

)

13

and, to avoid unmotivated responses, subjects were offered a lottery ticket for every correct prediction.

14

After the predictions of each player were registered, they were then invited to make their own strategic choice.

In the Colour treatments the computer informed players of the colour of their opponent at the beginning of each round. In the No-Colour treatments no information was given about one's opponent. When all subjects had registered their

predictions (of their current opponent's choice) **and** punched in their choice of strategy, the round was over and their screen would provide the following information:

- (i) His/her opponent's choice (and thus his/her payoff from this round)
- (ii) The group's aggregate behaviour in both the last round and for all rounds so far (on average); e.g. 30 per cent chose *h*, 60 per cent chose *d* and 10 per cent *c*

- (iii) The running total and the average of *his/her* payoffs for all rounds so far
- (iv) The average payoffs of the group for all rounds so far

In Colour sessions players were given additional information on:

- (v) The aggregate behaviour of all red players and of all blue players separately, both in the last round and for all rounds so far (on average)
- (vi) The running average payoff of blue and of red players separately

As is common practice in experiments of this type, the purpose of giving feedback to subjects in experiments is to remove sampling error and speed up convergence, thus avoiding the concentration lapses (not to mention spiralling costs) caused by a greater number of rounds. A printout of the screen offering a snapshot of what the players saw during the sessions can be found in

[Appendix 11.1](#)

11.4 Results

The theoretical predictions for the four treatments that come from standard and evolutionary game theory together with Rabin's fairness equilibria are summarised in [Table 11.3](#)

[Tables 11.4](#)

and

[11.5](#)

offer an overview of the experimental data. The data is expressed in percentages rounded-off to one decimal point. The data for the game that appear in **bolded** figures signify that the relevant observation in treatment **HD-HDC-Colour** is different from those in the same column (i.e. of the other treatments) at the 95 per cent confidence level.

¹⁷

The data here come from both the early, more 'noisy' rounds as well as the later ones (to which the predictions in

[Table 11.3](#)

apply more readily). Nevertheless there are three important results.

Result 1: Treatment **HD-HDC-Colour** stands out in terms of the frequency of the pure strategy Nash equilibrium (h, d) . In both games (HD and HDC) the frequency of the pure strategy Nash equilibrium (h, d) is significantly larger in this treatment than in the rest. In game HDC this difference becomes overwhelming.

¹⁸

Result 2: Behaviour in treatment **HDC-HD-Colour** is significantly distinct from that in **HD-HDC-Colour**, and rather similar to that in the No-Colour treatments. In particular, the frequency of outcome (h, d) in **HDC-HD-Colour** is statistically indistinguishable from the two No-Colour treatments (and, of course, significantly lower than in **HD-HDC-Colour**).

Result 3: Cooperative behaviour is present in the HDC game in all treatments.

[Table 11.3](#)

Predictions of long run, or equilibrium, behaviour

The predictions of conventional game theory

- (a) Behaviour will converge on one of the three Nash equilibria available: (h, d) , (d, h) or $Pr(h) = 1/3$

- (b) No prediction regarding order effects

The predictions of evolutionary game theory

<i>No-colour treatments</i>	
(a) One dimensional evolution will lead to the unique evolutionary equilibrium (also a Nash equilibrium in mixed strategies): $Pr(h) = 1/3$	(a) <i>Different colour meetings</i> : Two dimensional evolution leading to a unique evolutionary (pure strategy) equilibrium in which players holding one of the two colours play h and holders of the other colour play d
(b) The third strategy c will fade away in game HDC	(b) <i>Same colour meetings</i> : One dimensional evolution leading to the unique evolutionary equilibrium (also the Nash equilibrium in mixed strategies): $Pr(h) = 1/3$
(c) No prediction regarding order effects	(c) Strategy c will fade away in game HDC
	(d) No prediction regarding order effects

The predictions of Rabin's model of fairness

- 15
- (a) All cells on the diagonal of the payoff matrices of games HD and HDC may be observed systematically, in addition to the pure strategy Nash equilibria (h, d) and (d, h)
 - (b) Outcome (h, h) will occur more frequently in HD than in HDC (or, at least, not less frequently)
 - (c) Outcome (d, d) will occur more frequently in HDC than in HD (or, at least, not less frequently)
 - (d) The pure strategy Nash equilibria (h, d) and (d, h) will occur more frequently in HD than in HDC (or, at least, not less frequently)
 - (e) The use of strategy c in game HDC will not fade away
 - (f) No prediction regarding order effects
-

Table 11.4

Frequency (per cent) of outcomes in all 32 rounds of each game per treatment

<i>Game</i>	<i>HD</i>			<i>HDC</i>					
	<i>(h, h)</i>	<i>(h, d)¹⁶</i>	<i>(d, d)</i>	<i>(h, h)</i>	<i>(h, d)</i>	<i>(d, d)</i>	<i>(c, c)</i>	<i>(h, c)</i>	<i>(d, c)</i>
Treatment									
HD-HDC-NClr	29	39.8	31.2	36.7	9.8	3.7	6	30.2	13.6
HDC-HD-NClr	33	35.6	31.4	29.3	4.3	2	8.2	38.1	18.1
HD-HDC-Clr	21.4	51.8	26.8	19.2	38.7	2.2	9.3	20	10.6
HDC-HD-Clr	26.9	45.2	27.9	30.1	7.1	2.1	7.2	34.7	18.8

Result 1 is directly relevant to the *discrimination hypothesis* and is consistent with the two-dimensional evolutionary model.

19

This finding is reinforced by the data in

[Table 11.4](#)

showing that the more frequent occurrence of (h, d) in **HD-HDC-Colour** was achieved, especially in HDC, in spite of the fact that players did *not* play, in aggregate, h or d with frequencies significantly different to those in other treatments (see

[Table 11.5](#)

). It seems, therefore, that there was *something* in **HD-HDC-Colour** that enabled players to coordinate their h and d choices so as to boost the incidence of outcome (h, d) at the expense of (h, h) or (d, d) . (Whether this *something* was, in fact, the colour labels is the subject of our convergence analysis below.)

By contrast, **Result 2** goes beyond the two-dimensional evolutionary model as it

points to a clear order effect. Neither standard nor evolutionary theory can explain why the availability of strategy c from the outset seems to prevent the evolution of discrimination. Some emendation like our *sequence hypothesis* seems necessary.

20

Likewise **Result 3** is not predicted as an equilibrium outcome by standard or evolutionary game theory, but some care is required here as the play of c could result from errors or in the process of learning adaptively. The result is, however, consistent with Rabin's (1993) hypothesis (see

Table 11.3

). These options are considered in more detail in the next section.

To examine whether a discriminatory convention lies indeed behind the greater incidence of (h, d) in the **HD-HDC-Colour** treatment we use a version of Friedman's (1996) test for convergence. In each session I computed (separately for HD and HDC) the frequency p that, in cross-colour meetings, *blue plays h* and the frequency q that *red plays h* based on the last five rounds. If the null hypothesis that $p = q$ can be rejected, I proceed backwards to identify the round by which the discriminatory pattern observed in the last five rounds had settled down. Full details are given in

Appendix 11.2

, but the idea is to look for the largest number of rounds before the end which would give estimates of p and q which do not differ, (at a 95 per cent confidence level) from those values in the last five rounds. When $p > q$, then we say blue is advantaged (A) and red is disadvantaged (D) and vice versa.

The table in

Appendix 11.2

gives the results for whether convergence occurred in each of the sessions, which colour was advantaged by it, and by which round

convergence was achieved (if it was). In 15 of the 16 sessions of treatment **HD-HDC-Clr** convergence occurred within, on average, 15.9 (out of 32) rounds of HD. By contrast, only one of the eight sessions in the **HDC-HD-Clr** treatment showed convergence in the first part, the HDC game. Hence, there is evidence from the study of convergence which points to the emergence of a discriminatory convention in **HD-HDC-Clr** treatment; and there is evidence that availability of c from the outset prevented the evolution of a similar pattern in HDC in seven out of the eight **HDC-HD-Clr** sessions. In short, the sequence of play does appear to matter.

Table 11.5

Frequency (per cent) of strategies in all 32 rounds of each game per treatment

<i>Game</i>	<i>HD</i>		<i>HDC</i>		
<i>Strategies</i>	<i>h</i>	<i>d</i>	<i>h</i>	<i>d</i>	<i>c</i>
Treatment					
HD-HDC-NClr	48.9	51.1	56.7	15.4	27.9
HDC-HD-NClr	50.8	49.2	50.5	13.2	36.3
HD-HDC-Clr	47.3	52.7	48.5	26.9	24.7
HDC-HD-Clr	48.5	51.5	51	15.1	34

Table 11.6

summarises this evidence on the two hypotheses from data based on observations from all 32 rounds of each game. It shows:

- (a) that the null hypothesis based on standard game theory (i.e. that behaviour in **HD-HDC-Clr** is indistinguishable from **HD-HDC-NClr**) is rejected.
- (b) that the null hypothesis (i.e. that the sequence of play of HD and HDC makes no difference to behaviour in either game) is also rejected.

Instead, there is evidence that is consistent with the emergence of a discriminatory

convention based on colour identification in the colour treatment and evidence that the presence of **c** at the outset inhibits the emergence of a discriminatory convention.

In what follows, I focus on the 16 sessions where we have evidence that a discriminatory convention emerged well before the half point of the session (and regardless of which game was played first). Of these 16, 15 were sessions of the **HD-HDC-Clr** treatment and only one of the **HDC-HD-Clr** treatment (see

[Appendix 11.2](#)

for details).

[Table 11.7](#)

compiles data from these 16 sessions *from the last 11 rounds of HDC only*; that is, the reported frequency of outcomes emerged well after the convention had begun to take hold. Since the discriminatory conventions were well established by the time the

[Table 11.7](#)

dataset was compiled, we could identify whether each player was either advantaged (A) or disadvantaged (D) by the convention and so we plot the frequency of outcomes depending on whether the meeting is between mutually advantaged (A) or disadvantaged (D) players or between an advantaged (A) and a disadvantaged (D) player.

[Table 11.6](#)

Testing the hypotheses on aggregate data

23

	Null hypothesis; Alternative hypothesis in brackets	Sample sizes	p-values	
			HD	HDC
<i>Discrimination hypothesis</i> ²²	In HD-HDC-Clr, Freq of <i>h</i> by blue players = (\neq) Freq of <i>h</i> by red players	0560 choices by blue and 10560 by red players	0.04*	0.008*
	Freq of <i>h</i> in HD-HDC-Clr = (\neq) Freq of <i>h</i> in HD-HDC-NClr	10560 choices in HD-HDC-Clr and 2816 in HD-HDC-NClr	0.02**	0.009**
<i>Sequence hypothesis</i>	The proportion of sessions in which <i>Discrimination</i> evolved is (is not) identical across HD-HDC-Clr and HDC-HD-Clr	16 sessions of HD-HDC-Clr and 8 of HDC-HD-No Clr; discrimination was observed in 15 of the former and 1 of the latter	0.002**	

[Table 11.7](#)

reveals again the influence of convention. In colour sessions in which a convention did *not* become established

23

(see last row of

[Table 11.7](#)

), the pure strategy Nash equilibrium (*h*, *d*) occurs only 5.6 per cent of the time. In sharp contrast, in the sessions where a convention did emerge, we find that when A-players met D-players the pure Nash equilibrium of (*h*, *d*) is achieved with a very high frequency (81 per cent).

[Table 11.7](#)

also reveals another interesting difference. We find that in those sessions where the discriminatory convention emerged, there is a conspicuously high incidence (almost 90 per cent) of the cooperative (*c*, *c*) outcome between D-players. In comparison, there is mutual cooperation between A-players only 4 per cent of the time and there is a

negligible amount of cooperation between A and D-players. Likewise, when no convention emerges the level of mutual cooperation is strikingly lower at 8.2 per cent.

24

In other words, it seems that the part of the *sequence hypothesis* relating to cooperation receives support from the data in the sense that when a discriminatory convention emerges, it is associated with very high levels of cooperation between the D-players. The next section focuses on this result.

Table 11.7

Data from the last 11 rounds of game HDC

Mean outcome frequencies in the last 11 rounds of HDC in the 16 Colour sessions in which players of one colour (A) gained an advantage at the expense of players of the other colour (D)	(h, h)	(h, d)	(d, d)	(c, c)
Meetings between two A players	51.2%	8.9%	1.81%	4%
Meetings between two D players	2.1%	3.5%	0.5%	89.9%
Meetings between an A and a D player	8.2%	81%	0.4%	0.5%
Comparable frequencies in the 8 Colour sessions in which <i>no</i> discriminatory convention was established	22.1%	5.6%	3.1%	8.2%

Note: Bolded frequencies exceed the other frequencies in the same column with at least 99 per cent probability.

Table 11.8

Average payoffs per round (in Australian cents) of A-players and D-players in all 32 rounds of HD and HDC in treatment HD-HDC-Colour

	Game HD		Game HDC	
	A-players	D-players	A-players	D-players
Meetings between an A and a D player	66.3	21.7	137.8	39.7
Meetings between two A players	7.3	—	16.2	—
Meetings between two D players	—	19.6	—	101.3
Average	36.8	20.7	77	70.5

The combined influence of the discriminatory convention and this sequence effect can be seen from another angle in

Table 11.8

. This gives an analysis of the distribution of average payoffs. It shows that, over all rounds of treatment **HD-HDC-Clr**, A-players received 90 per cent of their money from meetings with D-players. On the other hand, 71.8 per cent of D-players' winnings came out of meetings with other D-players. Put differently, whereas only 5.8 per cent of A-players' earnings were due to cooperation with other A-players, D-players received 61.7 per cent of their total pay-out from cooperating with one another.

11.5 Why did cooperation occur among the 'disadvantaged'?

From the perspective of standard game theory, there seem to be two possible ways of explaining the high incidence of cooperation among D-players. One is to appeal to the heightened kind of rationality which can sustain cooperation among a subgroup by some version of punishment (or trigger) strategy. The difficulty with this interpretation is

twofold. First, there remains the question of why this is the only sub-group of D-players which managed to achieve cooperation in this way. Secondly, under any version of a punishment (or trigger strategy), when the game has a finite horizon players should abandon cooperation in the last round of the HDC game. However, the null hypothesis that the frequency of c play by D-players in the last round remains the same as that in the previous 31 rounds cannot be rejected at the 5 per cent level in favour of the alternative hypothesis that it fell (in fact it rose slightly).

The other way is to appeal to some kind of bounded or adapted rationality. Suppose, for instance, there is inertia with respect to strategy selection such that once an A-player learns to play h and the D-player learns to play d in cross-colour encounters, they unthinkingly do the same in same-colour matches. This would explain the high incidence of (h, h) among A-players but it would not explain why (c, c) results among D-players. Perhaps the D-players block the h strategy in mutual encounters (since they do not use it) so that they see a 2×2 version of the HDC which is a pure coordination game (see Bacharach, 1997, for a variable frame model of cooperation). In this coordination game, (c, c) could become focal on the basis of Pareto and risk dominance and, once established, it just becomes

the habit of D-players to play c with each other. The difficulty with this type of argument is that it presumes 'adaptive' players *unthinkingly* use particular strategies once they have been assigned to either the 'advantaged' or 'disadvantaged' role and the data casts some doubt on this.

25

Table 11.9

is drawn from the last one-third run (11 rounds) of treatment **HD-HDC-C1r** and gives the prediction-choice combination for both A-players and D-players. The first row reports that D-players predicted their opponent would choose c 879 times. In 861 out of those cases, they chose c themselves. A-players predicted c 789 times, but only responded with c in 31 cases (see row 3). In meetings with opponents bearing the same colour label as themselves, D-players cooperated almost every time they had predicted c (i.e. with frequency 98.7 per cent, see row 5). When they had not predicted c by a fellow D-player, they played c 43.3 per cent of the time (row 6). The latter is a high figure which provides some succour for the habit hypothesis, but since it is under half the figure for when they expected their fellow D-player to choose c , it seems that something more than a thoughtless attraction to c explains behaviour here. Likewise, although A-players in their mutual meetings are not attracted very often to play c , its frequency is higher when an A-player expects the other A-player to choose c (24.7 per cent, see row 7) compared with when they do not expect c (9 per cent, see row 8).

Likewise, rows 9 to 12 caution against this adaptive explanation. D-players seem to have thought quite carefully before attempting to cooperate. When they played against A-players whom they thought would *not* cooperate, they only chose c in two out of 722 cases (row 11); whereas, when they expected that the A-player would choose c , they cooperated in 31 out of 38 cases (row 9). Again this hardly accords with the view that D-players were thoughtlessly locked into playing c . Turning to A-players, their propensity to cooperate with a D-player was also influenced distinctly by whether they expected c or not (rows 10 and 12).

Since standard game theory does not seem able to provide convincing explanations of the persistence of cooperation (especially among D-players), we now turn to explanations which postulate psychological payoffs. The earlier discussion (

Section 11.2

) indicated how the Rabin model can explain cooperative behaviour in HDC and the conflict outcome (h, h) . Its drawback is that it cannot account for differences in the frequency of cooperative moves between our A-players and D-players. To make this possible, we would have to amend Rabin's model so as to explain why (c, c) is selected

as a fairness equilibrium among D-players but not among A-players.

One way of achieving this would be to assume that while playing HDC in treatment **HD-HDC-Clr**, agents' normative beliefs on entitlement reflect not just the structure of the payoff matrix (as Rabin, 1993, assumes) but, additionally, their role in the discriminatory convention which emerged in the earlier play of the HD game. So, A-players might have higher normative expectations regarding entitlements than D-players following the play of HD (see the average payoffs for the HD part of the game reported in

Table 11.8

). If this was the case, then (c, c) could be a 'fairness' equilibrium for D-players but not for A-players. Instead A-players with higher normative expectations may find themselves locked into a nasty (unkind)

fairness equilibrium with players of the same colour. In such an equilibrium they anticipate that their A-opponents are about to harm them by playing h and, in order to avoid the unfairness of repaying nastiness with kindness (or even with normatively neutral behaviour), they respond to a probable h with an h .

As suggested earlier, this kind of endogenous generation of entitlements follows a line of argument in Sugden (1986). It is not implausible, given what is known from other experiments (see Babcock *et al.*, 1995; Asdigan *et al.*, 1994; Schotter *et al.*, 1996; Binmore and Samuelson, 1993),

26

and it is a natural extension in some respects of what evolutionary theory suggests regarding the evolution of positive (i.e. predictive) beliefs into normative beliefs concerning entitlements. Our evidence seems to be adding to this line of thinking.

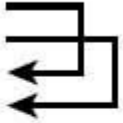
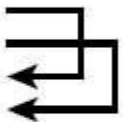


27

Nevertheless, the argument is, at best, suggestive. There are tricky issues of detail concerning precisely how entitlement norms evolve which need to be addressed. Furthermore, an appeal to the motivational force of an evolving set of psychological payoffs is not the only possible way to account for cooperative

behaviour among the D-players. For instance, it might be possible to argue that D-players 'group identify' and so adopt a form of team reasoning which produces cooperation (see Bacharach, 1999 who might explain this as a result of the 'common fate' hypothesis of group identity formation). The point of the argument in this section, then, is simply to lay the ground for a more thorough investigation along these lines because it seems that standard game theory cannot explain the cooperative behaviour among D-players while some kind of evolving fairness equilibrium or evolving group identification process could.

Table 11.9

The prediction-choice combinations of subjects in the last one-third-run (11 rounds) of HDC in treatment HD-HDC-colour

	<i>Player's colour</i>	<i>Opponent's colour</i>	<i>Player predicted opponent would play strategy</i>	<i>AND then played strategy</i>	<i>Conditional Freq (*)</i>	<i>%</i>	<i>p-values < 0.001(**)</i>
1	D	Any	<i>c</i>	<i>c</i>	861/879	98	
2	D	Any	$\sim c$	<i>c</i>	31/789	3.9	
3	A	Any	<i>c</i>	<i>c</i>	59/1174	5	
4	A	Any	$\sim c$	<i>c</i>	82/2603	3.2	
5	D	D	<i>c</i>	<i>c</i>	830/841	98.7	
6	D	D	$\sim c$	<i>c</i>	29/67	43.3	
7	A	A	<i>c</i>	<i>c</i>	20/81	24.7	
8	A	A	$\sim c$	<i>c</i>	74/826	9	
9	D	A	<i>c</i>	<i>c</i>	31/38	81.6	
10	A	D	<i>c</i>	<i>c</i>	39/1093	3.6	
11	D	A	$\sim c$	<i>c</i>	2/722	0.2	
12	A	D	$\sim c$	<i>c</i>	8/1777	0.5	
13	D	D	<i>h</i>	<i>h</i>	53/67	79	
14	A	A	<i>h</i>	<i>h</i>	602/826	72.9	
15	A	D	<i>h</i>	<i>h</i>	23/140	16.4	
16	D	A	<i>h</i>	<i>h</i>	28/1762	1.6	

(*)This column refers to the frequency of particular *combinations* of expectations and choices. For example, the first row reports that, in the last 11 rounds of HDC, there were 879 occasions when D-players predicted that their opponent would play *c*. Of those 879 instances, D-players decided to respond to that prediction by playing *c* 861 times. The sixth row reports that there were 67 occasions when, in a meeting between two D-players, a D-player did not predict *c* but played *c* regardless 29 (out of those 67) times.

(**)The *p*-values indicated here by the arrows relate to the null that the two frequencies linked by the arrows are equal.

11.6 Epilogue

This chapter reported on an experiment with two striking patterns of behaviour: the quick emergence of a relation of dominance in a repeated hawk–dove game associated with purely conventional labels; and a tendency for the subjects with subservient labels to cooperate with each other.

The first of these bears out the predictions of evolutionary game theory. The second cannot be explained by either standard or evolutionary game theory or Rabin's psychological theory. One possible explanation, however, comes from an amended version of Rabin's (1993) model: If the convention of dominance establishes a norm of different entitlements for those with different labels, then this norm could define a 'fairness' equilibrium among those with a subservient label which involves mutual cooperation. With this interpretation of the matter, the experimental data not only supports the hypothesis that 'psychological payoffs' matters but also that they are affected by the presence of a discriminatory convention. This is an important result, not least because it throws new light on the Athenian generals' argument, as well as Aristotle's famous maxim about the weak resorting to moral behaviour, with which I began the previous chapter.

VERDICT: The experiment presented here points unambiguously to two empirical findings of note: First, that discrimination based on utterly arbitrary characteristics evolves quickly and systematically in the experimental laboratory. Secondly, that game theory (of either the standard or evolutionary varieties) cannot explain this.

Utilising standard tools invented by neoclassical economists and adopting

experimental methods of the highest standards – as demanded by neoclassicists – we showed that people behave in a manner that neoclassical economists cannot explain. Moreover, we showed that these ‘unexplained’ behavioural patterns are highly significant as they hold the key to biases that we observe daily in society’s distribution of income, wealth, privileges, as well as a variety of social roles.

We also showed that these results, while unfathomable to the neoclassical mindset, have perfectly good explanations if one is prepared to look beyond neoclassicism.

Finally, the said experiment was published in *The Economic Journal* after having passed all the refereeing tests, checks and balances that are part and parcel of neoclassical economics’ strictures.

‘And to what effect?’ one might ask? What was the response of the neoclassical profession? Did any of its proponents feel the need to offer a rejoinder? To question our method? To carry out some other experiment whose results might cast doubt on our claims? No, dear reader. Silence. The paper might as well never have been published.

28

Only in economics is it possible that a powerful discipline’s basic tenets are disputed in one of its prestigious, mainstream journals but the ‘profession’ proceeds as if nothing has happened. No further evidence is needed that neoclassical economics is a kind of theocracy hiding behind equations and statistical methods but bent on remaining unperturbed by scientific inquiry and inconvenient facts.

Appendix 11.1: The 32 sessions of the four treatments

Abbreviations of the four treatments.

<i>Treatment</i>	<i>1st Game (32 rounds)</i>	<i>2nd Game (32 rounds)</i>	<i>Colour labels assigned?</i>
HD-HDC-NClr	HD	HDC	No
HDC-HD-NClr	HDC	HD	No
HD-HDC-Clr	HD	HDC	Yes
HDC-HD-Clr	HDC	HD	Yes

In each treatment subjects played the first game 32 times and then played the second game another 32 times. Below the sessions are listed in chronological order. Column N denotes the number of subjects in each session.

	<i>Treatment</i>	<i>N</i>		<i>Treatment</i>	<i>N</i>		<i>Treatment</i>	<i>N</i>
1	HD-HDC-NClr	24	12	HD-HDC-Clr	18	23	HD-HDC-Clr	18
2	HDC-HD-NClr	16	13	HDC-HD-Clr	18	24	HD-HDC-Clr	16
3	HDC-HD-NClr	22	14	HDC-HD-Clr	20	25	HD-HDC-Clr	22
4	HDC-HD-NClr	22	15	HD-HDC-Clr	18	26	HDC-HD-Clr	18
5	HD-HDC-NClr	18	16	HD-HDC-Clr	16	27	HD-HDC-Clr	22
6	HD-HDC-NClr	24	17	HDC-HD-Clr	20	28	HD-HDC-Clr	16
7	HD-HDC-NClr	22	18	HDC-HD-Clr	24	29	HD-HDC-Clr	26
8	HDC-HD-NClr	16	19	HD-HDC-Clr	24	30	HD-HDC-Clr	18
9	HDC-HD-Clr	18	20	HD-HDC-Clr	16	31	HD-HDC-Clr	22
10	HD-HDC-Clr	16	21	HD-HDC-Clr	20	32	HD-HDC-Clr	26
11	HD-HDC-Clr	26	22	HD-HDC-Clr	16			

<i>Treatment</i>	<i>No. of sessions</i>	<i>No. of players</i>	<i>Interactions per game</i>
HD-HDC-NClr	4	88	1408
HDC-HD-NClr	4	76	1216
HD-HDC-Clr	16	330	5280
HDC-HD-Clr	8	146	2336
<i>Total</i>	<i>32</i>	<i>640</i>	<i>10240</i>

discriminatory pattern which was observed over the last 5 rounds had settled down.

STEP 2: Following Friedman (1996), the following convergence criterion was used:

$$\frac{1}{L} \sum_{t \in R} \text{SupNorm}\{\pi' - \pi, \vartheta' - \vartheta\} \leq \varepsilon$$

where L is the length of run R under scrutiny.

Values π and θ , as before, were computed over the last five rounds of the game in question. Values π' and θ' were computed over the run of length L . At first we set $L = 6$ and chose as our 6 observations the last 6 rounds of the game. Thus, run R initially included the last 6 rounds of each game in each session. If the criterion was met for the chosen value of ε (see below for an explanation of how ε was chosen), L was set equal to 7 (i.e. R became the last 7 rounds of the game) and the criterion was computed again. This process ended at $L = \lambda - 1$ when the criterion was, for the first time, not met (given the same value of ε). At that point the algorithm came to an halt and convergence to a stable pattern of discrimination was pronounced to have occurred on round $32 - \lambda$.

The meaning of the above criterion is that the larger absolute deviation between (a) the empirical probabilities over the run's L rounds that A-players and D-players will play strategy h , and (b) the same empirical probabilities as observed *in the last 5 rounds*, the smaller the chances that the pattern of discrimination which we observe in the last 5 rounds had 'settled down' L rounds before the game's end. Thus, the criterion checks that the larger absolute deviation between (a) and (b) must **not** exceed a certain threshold ε .

Finally, the value of ε was selected in such a manner that if the convergence criterion were to hold then we could be certain with 95 per cent confidence that, in the last L rounds of the game, π' and θ' had converged to their values in the last 5 rounds. The table below, based on the above algorithm, reports on whether convergence was achieved and if so during which round:

(2) *Convergence table*

Session no. and colour treatment		Game HD			Game HDC		
		Conver- gence?	Which Colour?	Which Round?	Conver- gence?	Which colour?	Which Round?
9	HDC-HD	Yes	Red	26	No	—	—
10	HD-HDC	Yes	Red	24	Yes	Red	12
11	HD-HDC	Yes	Blue	19	Yes	Blue	8
12	HD-HDC	Yes	Blue	18	Yes	Blue	5
13	HDC-HD	Yes	Blue	1	Yes	Blue	26
14	HDC-HD	Yes	Red	24	No	—	—
15	HD-HDC	Yes	Red	21	Yes	Red	11
16	HD-HDC	Yes	Blue	20	Yes	Blue	2
17	HDC-HD	Yes	Red	23	No	—	—
18	HDC-HD	No	—	—	No	—	—
19	HD-HDC	Yes	Blue	15	Yes	Blue	8
20	HD-HDC	Yes	Blue	20	Yes	Blue	6
21	HDC-HD	No	—	—	No	—	—
22	HD-HDC	Yes	Red	14	Yes	Red	13
23	HD-HDC	Yes	Blue	16	Yes	Blue	1
24	HD-HDC	Yes	Red		Yes	Red	2
25	HD-HDC	Yes	Blue	10	Yes	Blue	20
26	HDC-HD	No	—	—	No	—	—
27	HD-HDC	Yes	Red	20	Yes	Red	21
28	HD-HDC	Yes	Red	7	Yes	Red	6
29	HDC-HD	No	—	—	No	—	—
30	HD-HDC	Yes	Blue	18	Yes	Blue	7
31	HD-HDC	No	—	—	No	—	—
32	HD-HDC	Yes	Blue	16	Yes	Blue	10

Appendix 11.3: Disaggregated data from all 32 rounds of treatment HD-HDC-Clr

In this appendix we present the data for *all* 32 rounds of *each* game in **HD-HDC-Clr** corresponding to

[Table 11.7](#)

(in which only data from the last 11 rounds of HDC was reported). Bolded figures signify that the relevant observation was different from those in the same column at the 99 per cent confidence level. Note that only data from 15 out of the 16 sessions of **HD-HDC-Clr** were used (since in session 31 – see

[Appendix 11.2](#)

– no colour emerged as ‘advantaged’).

Data from all 32 rounds of HD in the 15 *HD-HDC-Clr* sessions in which A and D colours emerged:

	Outcomes			Strategies	
	(h, h)	(h, d)	(d, d)	h	d
Meetings between two A players	30.9	44.2	24.9	53	47
Meetings between two D players	26.8	43.2	30	48.4	51.6
Meetings between an A and a D player	14	59.9	26.1	44	56

Data from all 32 rounds of HDC in the same 15 *HD-HDC-Clr* sessions as above:

	<i>Outcomes</i>						<i>Strategies</i>		
	<i>(h, h)</i>	<i>(h, d)</i>	<i>(d, d)</i>	<i>(c, c)</i>	<i>(h, c)</i>	<i>(d, c)</i>	<i>h</i>	<i>d</i>	<i>c</i>
Meetings between two A players	42.8	17.2	3	3	24.2	9.8	63.5	16.5	20
Meetings between two D players	17.7	12.8	6	29	10.3	24.3	29.2	24.5	46.3
Meetings between an A and a D player	8.2	62.3	0	2.6	22.7	4.2	50.7	37.3	16.1

The next table presents a further breakdown of the above data as it pertains to meetings between an A and a D player. Note that the data refers to game HDC (with the corresponding data from game HD in brackets). For example, in HDC there were no occurrences of *(d, d)* when an A-player met a D-player whereas that outcome occurred 26.1 per cent of the time when an A-player met a D-player in game HD.

Aggregate behaviour in HDC (HD data in brackets) when an A-player met a D-player in HD-HDC-Clr:

	<i>D-player</i>				<i>Sub-total</i>
		<i>h</i>	<i>d</i>	<i>c</i>	
A-player	<i>h</i>	8.2 (14)	62.3 (48.1)	8.5	79 (62.1)
	<i>d</i>	0 (11.8)	0 (26.1)	1.2	1.2 (37.9)
	<i>c</i>	14.2	3	2.6	19.8
	<i>Sub-total</i>	22.4 (25.8)	65.3 (74.2)	12.3	100

Appendix 11.4: Disaggregated data from all 32 rounds of treatment HDC-HD-Clr

This appendix offers three tables equivalent to those of

[Appendix 11.3](#)

only this time for treatment **HDC-HD-Clr**. Bolded figures again signify that the relevant observation was different from those in the same column at the 99 per cent confidence level. As in

[Appendix 11.3](#)

note that only data from the sessions of **HDC-HD-Clr** in which discrimination on the basis of colour emerged were used. That is, the data below refers to only four out of the eight sessions of treatment **HDC-HD-Clr** for game HD and only one session for game HDC (see

[Appendix 11.2](#)

).

Data from all 32 rounds of HD in the 4 **HDC-HD-Clr** sessions in which A and D colours emerged:

	<i>Outcomes</i>			<i>Strategies</i>	
	<i>(h, h)</i>	<i>(h, d)</i>	<i>(d, d)</i>	<i>h</i>	<i>d</i>
Meetings between two A players	31.7	39.9	28.4	51.6	48.4
Meetings between two D players	31.9	36.1	32	49.9	50.1
Meetings between an A and a D player	21.8	52.4	25.8	48	52

Data from all 32 rounds of HDC in the single **HDC-HD-Clr** session where discrimination surfaced:

	<i>Outcomes</i>						<i>Strategies</i>		
	<i>(h, h)</i>	<i>(h, d)</i>	<i>(d, d)</i>	<i>(c, c)</i>	<i>(h, c)</i>	<i>(d, c)</i>	<i>h</i>	<i>d</i>	<i>c</i>
Meetings between two A players	32.5	8.9	0.4	7	35.3	15.9	54.6	12.8	32.6
Meetings between two D players	32.4	7.2	4.3	10.1	30	16	51	15.9	33.1
Meetings between an A and a D player	27.8	6.2	1.8	5.8	36.8	21.6	49.3	15.7	35

Aggregate behaviour in HDC (HD data in brackets) when an A-player met a D-player in HD-HDC-Clr:

		<i>D-player</i>			
		<i>h</i>	<i>d</i>	<i>c</i>	<i>Sub-total</i>
A-player	<i>h</i>	27.8 (21.8)	3 (31.9)	17	47.8 (53.7)
	<i>d</i>	3.2 (20.5)	1.8 (25.8)	11.5	16.5 (46.3)
	<i>c</i>	19.8	10.1	5.8	35.7
	<i>Sub-total</i>	50.8 (42.3)	14.9 (57.7)	34.3	100

Appendix 11.5: Payoffs

(1) Overall average payoffs per player per round:

	<i>HD-HDC-NClr</i>	<i>HDC-HD-NClr</i>	<i>HD-HDC-Clr</i>	<i>HDC-HD-Clr</i>
HD	13c	1c	35.8c	19.3c
HDC	48.7c	29.5c	93.3c	74.7c

(2) Payoffs per player per round in colour sessions where discrimination evolved

<i>Treatment</i>		<i>HD-HDC-Clr</i>		<i>HDC-HD-NClr</i>	
<i>Pairing</i>	<i>Game</i>	<i>A's payoff</i>	<i>D's payoff</i>	<i>A's payoff</i>	<i>D's payoff</i>
Meetings between an A and a D player	HD	88.3c	15.7c	48.6c	69.2c
	HDC	\$1.36	39.7c	\$1.93	\$1.36
Meetings between two A players	HD	7.3c	—	\$1.18	—
	HDC	16.2c	—	\$2.43	—
Meetings between two D players	HD	—	19.6c	—	\$1.3
	HDC	—	\$1.01	—	\$2.45
Mean per game	HD	47.8c	17.7c	83.1c	\$1
	HDC	75.9c	70.5c	\$2.18	\$1.91
<i>Overall Average</i>		<i>61.9c</i>	<i>44.1c</i>	<i>\$1.51</i>	<i>\$1.45</i>

Notes

- 1 This chapter reproduces, to a large extent, Hargreaves-Heap and Varoufakis (2002).
- 2 Note that this third strategy was never related to subjects as 'cooperative'. Strategies were only referred to by their number.
- 3 Since individual behaviour is unobservable, and there is no room for trigger strategies to develop due to replacement of one's opponents after each round, players cannot invest in some reputation. Thus, each round resembles a one shot game.
- 4 Our choice of colours is not random. Mehta *et al.* (1994) report on a laboratory experiment of the 'name any colour' type which shows that blue and red are, roughly, equally salient. This is important because we wanted to preclude an additional source of salience; e.g. a situation in which *at the very outset* players of one colour (i.e. the one with higher salience) are seen as more likely to play aggressively as those of the other (i.e. the less salient) colour.
- 5 The opposite of course would be true if, at the outset, the frequency of 'h' among the 'reds' were to fall below both 1/3 and that of the 'blues'.
- 6 'The intuition is that a stable mixture of hawks and doves will evolve in a single population, but with two interacting populations, one will become all hawks and the other all doves.' Friedman (1996), p. 7.
- 7 It is worth remarking that there are other possible explanations for the emergence of such a convention. For instance, it might be explained by a version of Variable Frame Theory (see Bacharach and Bernasconi, 1997).
- 8 Of course the darker side of Rabin's (1993) fairness model is that Cressida may also value outcome (h, h) if she thinks that Troilus played 'h' not because he anticipated 'd' from Cressida, but because he expects her to play 'h' and thus wants to hurt her. Then Cressida may derive more utility from (h, h) than from (d, h)! In equilibrium, (h, h) is sustained by the mutual pleasure of hurting each other.
- 9 A minimum payment of A\$10 was guaranteed. However this floor was binding in only four out of 640 cases.
- 10 As is conventional in the literature, anonymity coupled with random matching and the knowledge that one would never play against the same player twice prevents the game from becoming a repeated game and, instead, renders it evolutionary (in the sense that players on the one hand cannot deploy trigger strategies – which require that the *same* players play repeatedly against one another and strive to build a reputation on eponymity – while, on the other hand, they condition their behaviour to the group's aggregate trends). In fact the software used a simple algorithm to match players (which of course the players were unaware of). To ensure that in the 'colour' sessions all red players would be matched against a blue player an equal number of times (and vice versa), and that the matching protocol would be as close to random (which is what subjects were 'promised' it would be) as possible, the algorithm produced *per player* an equal number of pairings with a player of the same colour as of the opposite one. In aggregate, the algorithm guaranteed that in the 32 rounds of each game (HD and HDC) the distribution of blue-blue, red-red and blue-red pairs would be 1/4, 1/4 and 1/2 respectively.
- 11 The familiarisation rounds involved the first game of the session (that is, HD in treatments HD-HDC or the HDC game in treatments HDC-HD). Afterwards the computer checked, via two multiple choice questions, whether the players understood the way in which their payoffs would be decided. The session did not begin unless all subjects passed this mini-test.

12

In a separate set of experiments with a battle-of-the-sexes type of game, we have found that whether a player chooses among the columns or the rows can evolve into a powerful discriminant. For instance we discovered that in the standard 2X2 version of that game, there was a strong tendency towards the Evolutionary/Nash equilibrium which favours the row players. See Varoufakis (1996).

13

There are two ways for soliciting expectations about discrete events. One is to ask agents (as we did here) to predict which of the two (three) strategies his/her opponent would choose in HD (HDC). The second way is to invite them to tell us the odds, as they see them. The latter has the advantage of revealing more about the agents' subjective *p.d.f.* However it suffers from two disadvantages. One is the (usually mistaken) presumption that subjects are familiar with distributions (and that they can express accurately their beliefs in probabilistic terms). The second disadvantage is that, unlike the former technique, it makes it hard to devise a simple reward scheme which will motivate subjects to reveal their expected distribution accurately. In selecting the former we decided to opt for the simplest question (i.e. which strategy, 'h', 'd', or 'c' do you think is more likely that your opponent will choose?), the simplest payoff-structure (i.e. if your guess is correct you will increase your chance of winning a prize) and the simplest (to interpret) reply. Since the sample size was large, and the objective was to monitor the *trend* of changes in such predictions (as opposed to their mean and standard deviation), the advantages of discrete predictions were deemed considerable.

14

The lottery scheme was calibrated in such a way that if one predicted correctly all 64 choices by one's opponents, one would gain a 100 per cent chance of winning A\$10 *in addition* to the payoffs from the games.

15

The predictions below are derived from Rabin's (1993) model. In brief, if v denotes the marginal importance of money relative to the psychological payoffs, it transpires that the influence of the psychological payoffs is a diminishing function of v . For v values below certain thresholds, the diagonal elements of the payoff matrices become equilibria while the original Nash equilibria (h, d) and (d, h) drop out. Predictions (a) to (e) are based on the implicit hypothesis (consistent with Rabin) that there exists a random (exogenous) distribution of the v 's amongst our subjects. Naturally, as there are multiple equilibria and no theory of equilibrium selection, these predictions are based on the presumption that the likelihood of each equilibrium is proportional to the range of v values which supports it.

16

Due to the games' symmetry and the fact that all players were choosing among the rows, outcomes off the diagonal are reported as one: e.g. (h, d) data reports on the frequency of both (h, d) and (d, h) etc.

17

The statistical tests used here need to be qualified. Although common in the experimental literature, they are open to the criticism that they treat as independent what might, after all, be repetitions of a single (or a few) observation(s). (Nb. this would be indeed true if players converge quickly to a fixed response.) Nonetheless, such criticism is pertinent when the reported statistical significance is marginal. In cases, like ours, where the differences between treatments are large, there is no cause for concern.

18

In fact the frequency of (h, d) in game HDC of **HD-HDC-Clr** is four times greater than the second highest frequency of the remaining treatments. The null hypothesis that the frequency of this pure strategy Nash equilibrium is the same across treatments **HD-HDC-NClr**, **HDC-HD-NClr** and **HDC-HD-Clr** cannot be rejected at the 5 per cent level in either game HD or HDC. By contrast, the null that the frequency of outcome (h, d) in treatment **HD-HDC-Clr** is the same with that in the other three treatments is rejected for HD at the 5 per cent level and for HDC at the 1 per cent level.

19

Note that the observations from the No-Colour treatments are fully consistent with those reported elsewhere viz. one dimensional HD play (see, for instance, McDaniel, Rustrom and Williams, 1994).

20

Perhaps the availability of strategy 'c' does not derail the evolution of discrimination but, instead, slows it down. Indeed it is possible to show in the context of an evolutionary analysis of HDC that there exist trajectories which, initially, take the evolutionary process away from the equilibrium (e.g. by boosting the frequency of cooperative play) before returning to it.

21

Note that this test of our *Discrimination Hypothesis* is considerably biased in favour of the null hypothesis: The data used contains not only the early rounds (during which a fledgling convention had had no time to emerge) but also the same-colour meetings in which the *Discrimination Hypothesis* does not predict differences in 'h'-play between the red and the blue players. And yet despite of all this 'noise' which ought to have made it harder to reject the null, in treatment **HD-HDC-Clr** (see Table 6) the null was rejected handsomely.

22

p-values: The reported *p*-values refer to the empirical probability that the value of the relevant test statistic is as extreme or more extreme than its observed value assuming the null hypothesis to be true. For example, the *p*-value of 0.002 reported for the **Sequence Hypothesis** means that the null of order-independence in the colour sessions can be rejected with 99.8 per cent confidence.

Test statistics: Two pooled *t*-test statistics were used in connection to the **Discrimination** hypothesis. One tested whether the frequencies with which the blue and the red players chose strategy 'h' in *each session* of the **HD-HDC-Clr** treatment were equal; see the *p*-value marked with (□). The other compared the frequencies of 'h' in **HD-HDC-Clr** with that in **HD-HDC-NClr**; the relevant *p*-value is marked with (□□). The *p*-value viz. the **Sequence Hypothesis** is based on a simple two sample pooled *t*-test.

23

Largely because of the availability of c from the outset.

24

The hypothesis that D-players are more cooperative than A-players is even supported by the aggregate, noisy data (i.e. data from all 32 rounds of HDC). The table below demonstrates this:

<i>Null hypotheses</i>	<i>p-values; sample sizes in brackets</i>
$\text{Fr}(c [A, A]) = (<) \text{Fr}(c [D, D])$	0.04 ♦ (4928)
$\text{Fr}(c [D, D]) \rightarrow a \text{ and } \text{Fr}(c [A, A]) \rightarrow b$ where $a = (>)0$ and $b = (>)0$	0.000 ♦♦ (4928)

where $\text{Fr}(c|[A, A])$ is the frequency with which strategy c was chosen in meetings between two A-players etc. The p -values are underpinned by a similar pooled t -statistic which tests the null that the frequency of strategy c is the same in A-player meetings compared to D-player meetings (the relevant p -value is marked by ♦) and that the frequency of successful cooperation among D-players or among A-players vanishes (the relevant p -value is marked by ♦♦). [Note that a Wilcoxon non-parametric test, not reported here, gave similar results.]

25

Notice that such inertia is irrational. Instrumentally rational players (i.e. those capable of maximising their own payoffs given their information) would follow an emerging convention only in cross-colour matches. Why? Because in the absence of any guarantees of consistently aligned beliefs, the discriminatory convention offers them useful information about their opponent's likely beliefs and actions. However, in same-colour matches they are useless. Therefore, only by mistake will payoff maximisers allow habits which took shape in cross-colour meetings to spread into same-colour ones. Such inertia, or reinforcement, presumes that players pay no attention to the outcomes of strategies that they did not choose. For an interesting discussion see Erev and Roth (1998) and Erev, Bereby-Meyer and Roth (1999).

26

It is not unusual for players belonging to different groups to entertain different perceptions of fairness. For instance, commenting on the data from their dispute-resolution experiment, Babcock, Lowenstein and Issachoroff (1995) conclude thus: 'Even when the parties have the same information they will come to different conclusions about what a fair settlement would be and base their predictions of judicial behaviour on their own views of what is fair.' Asdigan, Cohn and Blum (1994) report the well known fact that men and women rationalise by means of different principles of distributive justice their different socio-economic status as well as that of others. See also Kahn, O'Leary, Krulowitz and Lamm (1980), Major and Adams (1983), and Major, Bylsma and Cozzarelli (1989). Schotter, Weiss and Zapater (1996) suggest, in effect, that such ideas of fairness may be endogenously generated. In an ultimatum game experiment involving 8 pairs of players, the 4 proposers who gained most money (out of the 8 proposers in each session) were given the opportunity to play again (against another responder). In these sessions the responders (who knew that the proposers were competing against each other) accepted, on average, lower offers than in sessions where the proposers did not compete. Thus it seems that players are prepared to accept a lesser position if there is some rationale for it. Likewise Binmore and Samuelson (1993) report that, in the context of ultimatum games, the normative expectations of responders and proposers change at different speeds due to the fact that the former have less to lose from rejecting unfair offers by the latter.

27

Our data on subjects' point estimates of their opponent's choice, though not presented here due to space restrictions, shows unequivocally that, as convergence to the discriminatory convention was approaching, our players predicted the observed behavioural patterns rather accurately. For example, D-players (A-players) increasingly predicted a higher (lower) frequency of h if their opponent was of the opposite colour. D-players anticipated a higher (lower) degree of cooperativeness from opponents of the same colour than A-players.

28

A most astounding case of the $1 \rightarrow 2 \rightarrow 1$ path in the diagram of the dance of the meta-axioms (see [Chapter 1](#)).

12 On the distinction between evolution and history

The impossibility of modelling behavioural mutations amongst political animals

12.1 Prologue

12.1.1 Background briefing

A degree in economics bears a striking resemblance with a course on still photography, without of course the aesthetic pleasure afforded by the latter. Indeed, studying economics at university translates into spending years studying ... 'stills'. Every single model in the microeconomics textbook is a series of snapshots of *homo economicus*, frozen in time, working out which of his choices, that do not violate some exogenous constraint, corresponds to maximum intertemporal utility.

The very use of the word 'intertemporal' suggests that time somehow manages to sneak in. Not so! *Homo economicus*' intertemporal choices are crushingly temporal. He chooses whether to save a dollar today by comparing the utility of what this dollar will buy him now with the 'current' utility occasioned by the thought of what he will be buying with that dollar, plus interest, next year. And on what basis does he make this comparison? But on the basis of his *current, given, sovereign, consistent* and *fully determining* set of preferences. *Homo economicus*, in this sense, can only make intertemporal choices if his future consumption of an apple is re-cast as a different 'contemporary' apple whose utility is then compared to the orange he is now holding in his hands. His future 'states' are thus transformed into current ones and his future utility is assumed to be known, and fully evaluated, now. Only by collapsing the future into the present, to form a seamless über-present, can microeconomics model intertemporal choices.

As for *change*, its study depends on comparing different snapshots on the basis of *hypothetical reasoning* the purpose of which is, paradoxically, to keep time ... still. So, the hapless student is shown a supply and demand diagram and is told that *if* the demand curve were to be elsewhere *then* the equilibrium price would also have been different. Note the crucial difference: The student is *not* told that if the demand curve *shifts* from one position to another then the equilibrium price will *respond* by *shifting* to a new level. Such a narrative would involve a dynamic analysis that the model cannot sustain; one that necessitates telling a story in real, historical time.

But as we know, both partial and general equilibrium neoclassical analysis breaks down, into an abyss of indeterminacy, the moment the clock is running.

This is why the good neoclassicist never uses phrases such as 'the demand curve shifts and, in response, equilibrium price moves to...'; for they know that 'shifts' and 'responses' happen in a universe where the clock is ticking continuously and in which their microeconomic analysis is out of its depth. Thus, they are punctilious in their use of hypothetical reasoning, employing a large number of 'if this, then that' statements, in order to narrate their price and quantity theories, while refraining from any phrases which could, in the court of intellectual honesty, be traced to some illegitimate real time, historical narrative.

The reader, at this point, may wonder whether the above is a little too harsh on neoclassicism. After all, there are countless dynamic models out there, populating a plethora of journals with stochastic differential equations capturing the movement through time of whole macroeconomies. This is so but, if we look more closely, each and every one of those models contain a single person (perhaps clones of that person too) or, equivalently, a single economic sector (or many sectors with precisely the same techniques of production). They are, in effect, Robinson Crusoe economies. In a recent

book,

1

my co-authors Joseph Halevi and Nicholas Theocarakis and I argued that economics in general, not just neoclassical economics, has serious trouble combining, into the same analysis, *complexity* and *time*. We argued that economists seeking a 'closed' model must either stick to a corn economy-like model (a single sector, single consumption-cum-capital good), and then adorn it with beautiful dynamics, or allow for dizzying complexity in the context of a static analysis. Both *complexity* and *time* cannot fit into the same 'closed' model. This could not, of course, be different for neoclassical economics. The result is this schizophrenia that we inflict on our students of a totally static microeconomics and a fully dynamised Robinson Crusoe-like macro-narrative which, of course, bears no resemblance to really existing capitalism.

The gist of the above is that economics is a deeply ahistorical discipline. Not by some accident but by conscious, careful, painstaking design. Ahistoricity was the price economists *chose* to pay in order to have a chance at achieving mathematical determinacy, the holy grail of their ilk. Their critics from different social sciences have been lambasting economics' ahistorical narratives for decades, but economics goes from strength to strength because, by effectively banishing history from their theoretical endeavours, they have created models which, via the dance of the meta-axioms, armed them with unassailable discursive power. The only chink in their armour is, of course, indeterminacy, which the third meta-axiom has been valiantly keeping under wraps (as this book has been illustrating in the preceding chapters).

In the midst of this intellectual drama, evolutionary game theory came to the fore with a remarkable claim: It is now possible, courtesy of its blend of evolutionary biology and game theory, to give an evolutionary flair to standard neoclassical accounts. If neoclassicism's foundations can be recast as Darwinian, leaving behind their original Leibnizian texture, all of a sudden mainstream economics might become historically relevant.

Some may protest that an evolutionary process differs from a historical one. But that would be nitpicking to a neoclassicist eager to escape from a static world into

a world full of flux and feedback effects linking outcomes with utilities, costs and concentration ratios, investment decisions and risky choices.

So, has evolutionary game theory, which is undoubtedly a neoclassical tool, finally allowed mainstream economics to claim that it is no longer necessarily ahistorical? Or that, at the very least, it is no longer constitutionally static because it has become capable of incorporating evolution in its models? This is the question that started the inquiry behind the present chapter.

2

12.1.2 The rest of this chapter

Evolutionary ideas have a long history in economics (see Hodgson, 1993) and vice versa. Alfred Marshall warned in his *Principles of Economics* that '[t]he Mecca of the economist lies in economic biology rather than mechanical economic dynamics...' (1891; XIV). The opposite influence had been famously declared thirty years earlier by Charles Darwin who, in his introduction to *Origin of the Species*, acknowledged the impression Malthus' political economy had left upon his thought.

3

And yet economics developed and matured along different lines, gaining in prestige the more it distanced itself from Marshall's advice. Mechanism triumphed as economists invested all their energies in a calculus of preference whose purpose was to establish the conditions for some static equilibrium. Thus, evolutionary ideas remained on the margins of the discipline where they, nonetheless, continued to offer useful insights on the grand issues (e.g. Schumpeter's dynamic analysis of capitalism) that mainstream economics was overlooking, engaged as it was with the minutiae of tatonnement, i.e.

the process of 'inching' toward the equilibrium by grouping in the dark, also referred to as 'equilibrium selection'.

It took a remarkable rapprochement between evolutionary biology and mainstream economics (around the late 1970s) before evolutionary mechanisms were imported into the latter. This 'evolutionary turn' of economic theory was combined with, and mediated by, *game theory*. Game theory had already promised to revolutionise conventional economics but, by the mid 1970s, was losing much of its momentum. The reason? Its inability to tie down its own models without assumptions that demanded too much of human reason. At that point, evolutionary biology came to the rescue, offering *game theory* a way out: instead of pinning down solutions by complex reasoning, it suggested an evolutionary process that would select the most successful behaviour amongst the competing candidates. The combination of the economists' mindset and the evolutionary biologists' notion of dynamics led to the establishment of *evolutionary game theory* (EvGT).

Since then, economists have been eagerly pursuing two parallel tracks: First, they strive to figure out the extent to which the evolutionary approach confirms their earlier analytical efforts, which had hitherto relied on static analyses of hyper-rational choice. Secondly, they put EvGT to work in order to illuminate the institutions and histories of contemporary capitalism.

This chapter begins by welcoming the mainstream economists' newfound interest in the evolutionary dynamics of the institutions that shape our daily lives.

However, it asks: *Is EvGT sufficiently evolutionary?* Can it grasp the essence of the way in which institutions and individuals interact over time *in a social context!* Is it capable of illuminating the particularities of capitalist societies and of the manner in which they manufacture institutional patterns of social power and discrimination? In short, is contemporary history reducible to the evolutionary processes envisaged by EvGT?

To answer these questions,

[Section 12.2](#)

surveys the literature for relevant theoretical and empirical findings,

[Section 12.3](#)

uses EvGT in order to explain primitive accumulation and

[Section 12.4](#)

queries whether, and to what extent, EvGT can illuminate humanity's transition from primitive to capitalist accumulation. Then,

[Section 12.5](#)

asks the main question: Is there a profound difference between evolutionary and historical accounts? As my answer to this crucial question is affirmative,

[Section 12.6](#)

takes matter further by discussing what I refer to as the 'liberating power of history.' Finally,

[Section 12.7](#)

offers the customary chapter epilogue.

12.2 Some theoretical and experimental insights made possible by evolutionary game theory

The critical moment in the formation of EvGT was the 'infiltration' of *game theory* by the ideas of Maynard-Smith and Price (1974) and Dawkins (1976). In the world of insects and birds, biologists demonstrated (both mathematically and empirically) that hierarchies emerge on the basis of nothing more than *arbitrary* differences in appearance (e.g. whether a bird has blue or red stripes on its back or an ant has a white spot on its head). Differences that are clearly uncorrelated with the animal's physical

power, skill or any other personal characteristic, were suddenly shown to play a crucial role in determining its share of the 'spoils'. It took a small leap of the game theorist's imagination to see this approach's potential for constructing a theory of institutionalised discrimination within human society.

To make this point clearly, consider the hawk–dove game below, which featured prominently in the experiment I reported in

Chapter 11

. Players may choose to be aggressive (*h*) or cautious (*d*). Mutual aggression leads to symmetrical 'injury' and the *loss* of 2 units of evolutionary fitness, where the latter is meant to measure the expected number of offspring.

5

In contrast, mutual caution means that the spoils (brownie points in the evolutionary stakes) are shared, and so are the evolutionary brownie points. However, when one of the two opts for the aggressive stance, while the other behaves cautiously, the former gets all the spoils (e.g. the nest they are fighting over) leaving the latter with nothing.

Hawk–dove game

	<i>h</i>	<i>d</i>
<i>h</i>	-2, -2	2, 0
<i>d</i>	0, 2	1, 1

Suppose next that players are drawn from a large population and meet fresh opponents each time. Evolutionary biology startled game theorists with a brilliant methodological strategy: Instead of modelling explicitly the players' reasoning that leads them to their chosen strategy (as game theory was struggling to do for decades), biologists focused on the strategies themselves and studied the way these evolve in response to their relative 'success'. The evolutionary idea here is that players are, somehow, *programmed* to choose a strategy at any point in time but, and this is the rub, also that strategies get 'copied' by other agents in proportion to the payoffs they yield *relatively to average payoffs in the population*.

It is now straightforward to demonstrate that, if the population is utterly homogeneous (i.e. players are perfectly identical and thus indistinguishable from one another), there exists a unique *evolutionary equilibrium*: one third of players will be acting like hawks and the rest will be playing cautiously (like doves) [see

Appendix 12.1

]. It is also straightforward to demonstrate [see

Appendix 12.2

] that, if players carry a distinctive feature (e.g. some have green eyes while the rest have blue eyes), these features will play a significant role in determining overall behavioural patterns *even if they are arbitrary and denote nothing about the player's character* (e.g. her talent, aggression, intelligence).

More precisely, EvGT proves that, in an evolutionary equilibrium, all the players with one of the distinctive features will play aggressively against all the players with the other feature who will invariantly acquiesce (e.g. all the blue-eyed players will play *h* against all the green-eyed players who will, in turn, play *d* against blue-eyed players; *or vice versa*). The theory cannot predict which group will dominate (the blue-eyed or the green-eyed); only that *some* group will!

In summary, evolutionary biologists helped game theorists understand that:

- (a) extraneous characteristics can 'seed' conventions which advantage one type of individual relative to another (even if the difference across individuals is arbitrary), *and*
- (b) the resulting conventional discrimination is stable because of the reluctance of individuals disadvantaged by it to risk subverting them.

These two results, taken together, echo the suspicion that those at the margin of the economics profession always harboured: that *discrimination is the result of evolved institutions which distribute social power in ways that have little or nothing to do with personal characteristics, aptitude or application*. Indeed, (b) above resonates nicely with the view wildly held within the social sciences (e.g. de Tocqueville, Marx, Foucault etc.) that the secret of systematic 'oppression' lies in the mind of the 'oppressed' (rather than in the mechanisms of 'oppression' consciously devised by the 'oppressors').

Nevertheless one cannot be too careful when transferring ideas from the biological sciences to social theory. So the question becomes: Granted that the asymmetrical distribution of resources in the animal republic is often founded on utterly random and extraneous characteristics (Question 1) 'Does this result from evolutionary biology extend to human societies founded on primitive accumulation?'

and (Question 2) 'What, if any, is the implication of an affirmative answer to the previous question for more complex human societies in which the webs of social power are fashioned largely in the realm of social relations of *production*?'

An affirmative answer to the first question was given in

Chapter 11

which reported on an experiment designed to test whether the results reported by the evolutionary biologists extend to humans. Experimental subjects were placed in a controlled environment where they played the hawk–dove game above, only this time with payoffs taking the form of dollars (rather than evolutionary brownie points).

To test EvGT's propositions (a) and (b) above in a human context, the experiment (as the reader who has read

Chapter 11

will recall) was run in two different formats: The first was the control treatment in which the games were played under conditions of complete anonymity. Subjects simply had no information whatsoever regarding their opponent/partner. Thus anonymity simulated an environment in which subjects cannot distinguish between their partners/opponents, thus rendering the population homogeneous at the level of individual perception. The second treatment tested propositions (a) and (b) directly by giving players a single piece of clearly extraneous information about their opponent/partner. What was this piece of information? And how are we so sure it was extraneous?

At the beginning of each of these sessions, players picked a card at random from a pile of cards half of which were blue and half red. Thus each player's 'colour' was determined. Once the sessions began, and in each round, subjects were informed of the colour of their partner/opponent. Naturally, this information was as extraneous as it could have been: everyone knew that, since it was commonly known that the colour assignment was random, it conveyed no significant information regarding their partner/opponent's character. The question then became: Would propositions (a) and (b) above be confirmed in this experimental setting?

What precisely would confirm it? It would be confirmed by the observation of significantly different degrees of aggression in cross-colour meetings between the 'blue' and the 'red' subjects. Which is precisely what was observed: In about ten rounds or so, one of the two colours had come to dominate the other. In other words, subjects of one of the two colours evolved a tendency to act more aggressively toward subjects of the other colour (than to subjects with the same colour as themselves). In some sessions it was the blue players that dominated the red while in others the reverse was observed. Moreover, subjects with the colour that evolved as 'inferior' developed a tendency to 'submit' to the enhanced aggression of their differently 'coloured' opponents by adopting a far more cautious approach to them (compared to the average incidence of cautious behaviour observed in experiments without any colour assignments).

In short, it seems that the biologists' results regarding the evolution of arbitrary discrimination among animals extend nicely to human behaviour (at least under laboratory conditions). Why is this significant? And what does it have to do with the anatomy of institutions and social power? Well, it must be significant to anyone striving to argue that the observed distribution of income and, more generally, social roles may not necessarily be predicated upon differences in human

capital, aptitude, application etc. Of course this is not to argue that all hierarchies reflect nothing but arbitrary differences in appearance. What it *does* show is that we cannot take it for granted that a systematic pattern of discrimination, according to which some group dominates other groups consistently, is a reliable indication that the dominant group is substantially different from the subservient ones (let alone 'better'). Is this not what the first-wave feminists were arguing for?

Indeed, if patterns of highly differential income distributions, and robust discriminatory conventions, emerge within 45 minutes in our experimental laboratory (on the back of a random colour assignment), it takes a grandiose leap of faith to assume that the hierarchies we observe around us, outside the laboratory, are somehow free of arbitrariness. When an insignificant characteristic (like a random colour assignment) can be at the heart of intense social stratification, what should we expect of emotively charged bodily differences (such as different reproductive systems, skin colour etc.)? Thus the importance of these results for feminists, anti-racists etc.

From the perspective of this chapter, the issue at hand is the more general lesson we can draw from

Chapter 11

regarding the evolution of institutions and the manner in which they disperse social power. Admittedly, our society is far more complex than the world of primitive accumulation to which bees and birds (or, indeed, our experimental subjects

) are confined. The question therefore is: *What is the implication of these results regarding the history of human relations first in the context of primitive accumulation and subsequently for the more complex world of agrarian societies and, ultimately, capitalist dynamics?*

12.3 On the emergence of institutions, conventions and social norms under primitive accumulation

Institutions are defined broadly here. They constitute *any* mapping from individual motivation to social outcomes which cannot be reduced to data on private preferences and constraints. Economists might find it useful to re-interpret the socio-economic equilibria spawned by multi-dimensional evolution as institutions which rely on *conventions*, in the sense of Lewis (1969). What sustains the practice of, say, red players conceding in *hawk-dove*, while blue players take the lot, is simply the players' forecast that this is what will happen. Such predictions become self-fulfilling because, once they are shared, no individual can profit by acting in a manner that contradicts them.

Of course, the opposite prediction is equally self-sustaining, i.e. all players expecting that the reds will dominate the blues *provided the population held this alternative set of predictions*. Thus behaviour at each of these (potential) evolutionary equilibria is *conventionally determined*. The evolutionary approach, quite naturally, differs in the specificities of its interpretation depending on whether the agents under study are members of the animal republic or the human race. In the case of ants and bees, adaptive behaviour is all evolutionary biologists require to explain the evolutionary dynamics; any talk of institutions or convention is superfluous. However, when the players are human, the evolution of behaviour

is underpinned by (and gives rise to) an evolving belief system which, in turn, is

equivalent to an institution sustained by conventions. When, for instance, red players predict that blue players will act aggressively against them, the *power of their prophesy* entrenches an asymmetrical institution which, effectively, yields 'property rights' to the blue players (since the latter sooner or later realise that they are best off acting aggressively to the reds as a result of the observed fact that the reds are more acquiescent).

Thus hierarchies are *instituted* in response to the structure of the interaction; as opposed to a mere reflection of the distribution of the individuals' attributes, features and talents. When the games people play in order to feed themselves and find suitable shelter have asymmetrical equilibria (e.g. the asymmetrical Nash equilibria of the game in our experiment) then, as long as people bear even the most irrelevant distinguishing mark or feature, social evolution will spawn inequitable institutions whose 'function' is to minimise conflict, on average, by discriminating ruthlessly in favour of one group and at the expense of others.

But how do rational agents accept the logic of such arbitrary social divisions? An interesting answer comes from unexpected quarters: David Hume's explanation of *how mere conventions annex virtue to themselves and thus become social norms, or norms of 'justice'*. Cast in modern terms, the idea is that a community's institutions become more resistant to 'mutations' when people not only expect others to behave in accordance to the established conventions but, also, feel that deviating from them is somewhat... *wrong*; kind of... *morally defective*. Hume insisted that we learn not only to *predict* that others will follow the established convention but, additionally, that we *expect of them* to do so. Indeed, when they fail to do so, many of us are often filled with moral indignity at behaviour 'prejudicial to human society'.

At that point, our predictions *vis-à-vis* others' behaviour have become *normative, or moral, expectations*. In Hume's (1888) own words, at some point of the evolutionary path, the 'is' and the 'will' become a 'must' or an 'ought': '...when of a sudden I am surprised to find, that instead of the usual copulations of propositions, *is* and *is not*, I meet with no proposition that is not connected with an *ought* or an *ought not*.' Sugden (1986, 1989) expands on this theme with the point that as conventions begin to impart 'moral' beliefs', they gather additional resistance to behavioural mutations. Put simply, a convention that makes us not only predict that we shall all adopt a certain behaviour but, also, that *we ought to*, is far less susceptible to mutations. And since robust conventions minimise conflict and enhance benefits *on average*, morality is an illusion *functional* to the individuals' petty interests. Moreover, when a ruthlessly discriminating convention emerges, people find it difficult to accept that the convention is in some sense arbitrary while also being so discriminatory. So people remove the resulting *cognitive dissonance* by finding, or inventing, additional principles that will justify the actual convention because it is 'just', 'fair' etc. When they succeed in this, the convention becomes more entrenched as both its beneficiaries *and those it discriminates against* are less likely to contravene it.

In contrast to Kant who thinks that 'the majesty of duty has nothing to do with the enjoyment of life' (1855), Hume's disciples (see also Binmore 1998) see morality as the reification of conventions whose *raison d'être* is to coordinate behaviours to some equilibrium devoid of waste and conflict. They also see norms of justice in the same light; namely, as conventions that imbue people with expectations of what is right, just, or wrong. *At the political level*, this conversion of predictions to ethical beliefs gives rise to the notion of the 'common good;' which is, in this account, another illusion brought on by the observation that convention-following brings greater average benefits (unequally of course). *At the level of the individual*, as in our laboratory for instance, we observe that dominant colour members showed vivid signs of moral outrage when an opponent with the 'other' colour acted aggressively toward them. And all that after sixty minutes of

laboratory games in which the stratification was based on a random colour assignment! Is it any wonder that, after centuries of discrimination, many women feel that men *deserve* the leading social roles? Or that most men in Papua New Guinea accept the moral superiority of white, male, American Protestant preachers? To sum up, three are the exciting aspects of this evolutionary theory of institutions:

(a) That institutions divide the population along rigid lines of stratification with some groups profiting at the expense of others and *independently of the powers or aptitude of its members*

(b) That as iniquitous institutions evolve within the context of primitive accumulation, the resulting divisions have a tendency to subdivide and multiply further, thus creating institutional discrimination within the major social strata they have generated at an earlier stage of the evolutionary process (see Hargreaves-Heap and Varoufakis, 2004, Section 6.3.3 of

[Chapter 6](#)

)

(c) That the evolved conventions of distributing assets and roles asymmetrically spread from one realm (or game) to another *by analogy* (Sugden 1986, 1989).

This last point deserves some elucidation: Primitive accumulation takes many forms. Hunter-gatherers operating cooperatively in order to catch large prey (e.g. stags) must develop resistance to the centrifugal forces of prisoner's-dilemma-like urges (as J.-J. Rousseau knew all too well) that are best kept in check by conventions for dividing the spoils around the camp fire. In other settings (e.g. areas where the prey migrates or the weather conditions change rapidly from one season to the next), hunter-gatherers must nurture nomadic conventions for both hunting and distributive purposes. In large areas with scarce, small prey (e.g. hare), hunter-gatherers are more likely to work alone. However, they are still likely to come up against one another and compete over the same prey or resource, not unlike our experimental subjects in the laboratory (or indeed the bees and insects in Dawkins, 1976; Maynard Smith and Price, 1974).

The socio-economic context of hunter-gathering just described is more complex than the primitive accumulation set up in our experiments at least in one important sense: the coexistence of cooperative and non-cooperative interactions. Sugden's (1986, 1989) point is that there is good cause for thinking that the institutions

which evolve in response to the non-cooperative games extend by analogy to the more cooperative settings. Indeed, some recent (hitherto) unpublished experiments (conducted by this author) show that, once conventions have taken hold in the context of non-cooperative interactions, they spread by analogy, inertia and mimicry to cooperative games. For instance, once a pattern of dominance is established in simple accumulative contests of the *hawk-dove* variety, it colonises the ensuing, more complex bargaining contexts.

From a historical perspective, the above undermines the 'romantic' view of primitive societies as 'states of nature' devoid of social institutions. EvGT arms us with sufficient confidence to hypothesise that social institutions, and hierarchies, had probably evolved even before we were 'fully' human. As geographical and climatological conditions necessitated more cooperative patterns of primitive accumulation (e.g. nomadic or collective hunting), these hierarchical conventions spread by analogy from the realm of *hawk-dove* like interactions to the ways and means by which collective produce was privately appropriated. To the extent that the community's evolutionary fitness was intimately linked with the solidity of those conventions, developments that weakened any tendencies to 'disobey' the established conventions were reinforced.

The evolutionary fitness of these institutions was improved further by two separate developments: First, the subdivision of populations into sub-strata entrenched the conventions of discrimination, by ensuring that a significant minority of the 'victims' of

'main' source of discrimination had a stake in preserving the overall pattern of discrimination as they derived *some* benefits from it when interacting with a small number of other substrata (see Hargreaves-Heap and Varoufakis, 2004, Section 6.3.3 of

Chapter 6

). Secondly, the evolution of human language, around 100,000 years ago, which facilitated, through the invention of moral signifiers, the emergence of concomitant ethical beliefs that 'enabled' people to feel not only that the violation of given conventions is dangerous but that, more poignantly, it is also *morally problematic* (recall Hume's ironic point of the ease with which our language slips from 'is' to 'ought' statements.)

As one might expect of a model whose roots are to be found in evolutionary biology, the preceding theory offers a full account of the birth of a great variety of coexisting institutions even within rather primitive forms of society. The question, however, is whether this type of analysis is adequate for explaining the variety of institutions observed in more complex societies, ranging from the agrarian to the capitalist.

12.4 From primitive accumulation to capitalism

Humanity's *Great Leap Forward* came with the development of farming which put us on the path of *socialised production* (a prerequisite for sustainable farming practices), *organised armies* (for the protection and/or appropriation of stockpiled food), *bureaucracies* (for the organisation of collective effort and the distribution of the resulting surplus), *writing* (for the purposes of book-keeping), the *evolution of differential resistance to new diseases* (leading to the genocide of those without it by those with it; e.g. Native Americans and Aboriginal Australians), the *technological progress* that led to greater capacities to create (e.g. metal technology for the manufacture of ploughs) as well as to destroy (technological advances in the development of weaponry) etc. However, even before we embarked collectively down that path, we came to it fully equipped with institutions founded upon the discriminating conventions developed at the earlier, hunting-gathering stage of socioeconomic development.

The hierarchical norms of dividing contemporary goods and chores did not begin with socialised food production. As the latter did not replace hunting-gathering abruptly, but coexisted with it for centuries (see Diamond, 1996), underneath the surface of the norms of surplus distribution there are many layers of prior discriminatory conventions which have their roots in an earlier hunting-gathering era. By simple deduction, the norms that determined who controlled the land fed into new, analogous norms regarding control of the surplus. Indeed in the previous section I argued that primitive accumulation leads inexorably to discriminating conventions by which the contested assets are distributed systematically in favour of one group and *for reasons that may have nothing to do with its members' personal characteristics*. If this is true about hare, stags, fruit and roots, it must be also true about other-assets such as fertile pieces of land. With a minor leap of the imagination we can visualise the conversion of relatively primitive distributive norms to complex norms of distributing: (a) the work load in the fields, warehouses, barracks etc., and (b) the share of the agricultural production enjoyed by each.

However, the moment food production comes into the picture, the epicentre of social power shifts from appropriation-*cum*-consumption to control over the production process. The simpler institutions of primitive accumulation can hardly carry the burden of this major socio-economic transformation. Rituals for dividing spoils and determining hierarchies around the camp fire are one thing; rules governing access to land, the division of labour between farmhands, smiths, priests and soldiers etc. are quite another. The emergence of agrarian economies in the midst of tribal life required a

different kind of institution capable of dispersing a new type of hitherto unknown social power: the power to control surplus production or, more briefly, *extractive power* (see C. B. McPherson, 1973, for the original articulation of this notion). Below I offer a re-worked definition of *extractive power*, one which takes on board some of EvGT's conclusions:

Generally speaking, person *i* exercises extractive power over *j* if:

- (a) *i* and *j* are virtually identical except that *i* sports extraneous feature *F* which places her in the advantaged social group A, leaving *j* in disadvantaged group *D*
- (b) *i* can persuade *j* to perform task *T* which results in surplus *S*
- (c) *i* can, courtesy of her membership, enforce property rights over *x per cent* of *S*
- (d) *j* would not have performed task *T* for $100 - x$ per cent of *S* had the distinction between group *D* and *A* not evolved previously
- (e) Social norms prevail upon *i* and *j* to think of the $[x, (100 - x)]$ per cent distribution as 'fair'

Extractive power is thus a straightforward extension of asymmetric conventions for distribution of non-produced goods to a community which produces assets in the context of collective manufacture. In principle, extractive power can emerge in hunter-gatherer communities too; in the sense that some group may develop, theoretically, a capacity to compel others to hunt/gather on their behalf. However, such conventions are less likely to take hold and command a significant proportion of work effort when individuals have the opportunity to abscond and fend for themselves. The more restrictive the access to productive resources (e.g. the more fences there are around fertile land) the greater the preponderance of extractive power.

In this account, social strata which gained conventional control over scarce land acquired also conventional control over others' productive efforts. Extractive power became, in this manner, inextricably linked to the technology of production and the *outside options* of individuals belonging to groups devoid of extractive power. The power to compel under the definition above is not the form of power associated with brute force but the subtler type of power which relies on making offers that the 'other' cannot refuse as a result of lacking viable outside options.

8

Moreover, the possibility of coexisting and inter-weaving patterns of extractive power allow for the possibility of older and newer conventions to operate side by side within the same institutional framework; at least for a while, blurring further the distinction between the dominant and subservient groups. The group privileged by history in the land-distribution game (e.g. the landed aristocracy) became a social class once (a) its privileges became hereditary (and the group could reproduce itself as a group), and (b) it embellished its extractive power over the rest with moral meaning (i.e. a dominant ideology). The latter was subsequently reinforced by the complexity of the conventions by which control of the land and its output was dispensed and the normative beliefs in which they were disguised. History books tell us that, when the exercise of extractive power became too obvious, revolt beckoned and, quite often, the heads of the dominant group's members rolled. In short, the greatest defence of the conventions of the first societies to produce surplus was the capacity of norms founded on extractive power to become invisible.

This capacity of distributional conventions to hide under multiple veils reached its pinnacle with capitalism. Spartacus became legendary because he personified the liberation of slaves from the normative beliefs that maintained a culture of quasi-voluntary submission to the naked extractive power of their Roman owners. However, his task was made considerably easier by the very nakedness of the extractive power that the slaves were subject to. Without self-ownership, and with the whip of the slave drivers swirling above their heads, slaves were ripe for the revolt that Spartacus drew them into. All that was necessary was a whiff

of optimism about the prospects of their rebellion. But when extractive power is maintained against a background of comprehensive negative liberty, the normative beliefs accompanying the conventions underpinning capitalist relations of production become considerably more oblique.

9

As already mentioned, farming introduced extractive power by shifting the centre of social life from the norms of distribution of exogenously generated assets to the norms of distributing land, labour and the resulting output. Capitalist production added another crucial complication to the 'story': extraction by property owners of the producers' output was shifted from the *post-* to the *pre-* production phase. Rather than collecting by stealth part of the output after it was produced (as was the feudal lords' wont), capitalists paid in advance a retainer for the workers' services; a retainer large enough to secure their surrender of *future* time and toil but less than the *expected* value of their labour.

Put simply, capitalism reversed the timing of extraction. Rather than receiving a fixed amount of the produced goods, the members of the socially dominant group would advance a fixed amount to the workers and claim the residual. Had we not experienced the momentous change that followed this reversal (i.e. the industrial revolution with all its wonders and catastrophes) perhaps it would not have been immediately obvious why it matters so much. After all, what does it matter who retains the residual? Paul Samuelson once famously claimed that who pays whom in the production process (the capitalists paying the workers or vice versa) should not matter. The reason it *does* matter is twofold.

The best rehearsed explanation is that, having laid out a fixed amount to the workers at the outset, capitalists acquire an incentive to squeeze as much produce out of them in the ensuing production process. A second explanation which receives little attention concerns the pivotal role this reversal played in disguising the social conventions at work. Under pre-capitalist social relations of production, control over production largely remained in the hands of the producers. It was only after the crop came in that the distributional conventions would kick in; a fact that made obvious the evolved and utterly arbitrary extractive power that the owners of land had over the non-owners. But under capitalism, the temporal reversal of residual claims meant that workers lost control over the production process. For the first time in human history the residual claimants paid in advance for the privilege of exercising their extractive power. Given the inherent risks of paying for something in advance, the task of removing the cognitive dissonance resulting from the preposterous social asymmetries that capitalism brought to the fore was eased substantially.

Those privileged by the new capitalist conventions could legitimise their gains based on the mythical notion of profit as a just reward for risk-taking. More importantly, those disadvantaged by the same conventions could live with their situation more easily by a combination of normative beliefs shaped by: (a) the seemingly symmetrical position of capital and labour ('we receive profit in return for laying out in advance our capital, and you receive this capital in advance in return for your labour'), and (b) the soothing impact of negative liberty for all.

The near-perfect invisibility of the social conventions at the heart of the institutions of capitalist production thus played a central role in solidifying the former and stabilising a system which proved remarkably successful at weathering all types of self-generated crises. From this paper's perspective what matters is the nexus between this invisibility and the varieties of coexisting patterned social power entrenched in contemporary institutions. In conclusion, the proliferation of coexisting institutions under contemporary capitalism may simply reflect simultaneously:

(a) the deepening antagonistic character of the games we play as technology makes it easier for telephonists in India to take emergency calls from Colorado and production to

be shifted at a moment's notice in search of the lowest wage rate

(b) the increasing fragmentation of the dominant ideology into post-modern, localised, ideologies that lack some 'common currency' (e.g. the demise of the Enlightenment ideals of liberal society) and, more importantly, make it hard for us to distinguish the overarching socio-economic system's structure

(c) the increasing tendency of capitalism to obfuscate the essence of distribution by altering the timing of payments and delivery of goods (i.e. the creation of futures markets that requires a great deal of technical expertise to disentangle)

(d) the ensuing crisis of the state whose authority is undermined both by (a) and (b) above.

12.5 Historical versus evolutionary approaches

In this section I shall argue that, despite its great and obvious merits, the evolutionary approach is severely limited when it comes to explaining history in general and capitalist history in particular. The notion of evolutionary equilibrium that social science has inherited from biology is too brittle to capture the subtleties and richness of human societies. Although it unquestionably brings many fascinating insights to the study of our species' historical and socio-economic dynamics, it is incapable of capturing some of its more poignant aspects; those very aspects that make human history what it is. Let me begin this argument by rehearsing, once more, the major insights of evolutionary theory. It demonstrates brilliantly how:

(a) conventions emerge depending on the shared salience of extraneous features of the way people hunt and gather, form beliefs, and learn from their interaction

(b) competition between conventions 'selects' some while condemning others to extinction depending on: (i) each potential convention's initial number of adherents, (ii) how they distribute the benefits of coordination across their followers, and (iii) their ability to skew interactions towards fellow users

¹⁰

(c) the institutions that correspond to these evolved, behavioural conventions cannot be undermined through subversive individual action; that only collective action can do this.

Now, let us take this last point about individual mutation *versus* collective 'revolt'. While biologists seem convinced that modelling mutations as random events does not jeopardise the predictive power of their theories, viz. the evolution of genes, the same presumption cannot be justified in social science. I am no biologist and thus I am prepared to accept the biologists' claim that it is scientifically unproblematic to model the mechanism which generates mutations in our genes as *statistically independent* of the mechanism that alters their individual functioning. But as a social theorist, I believe that a similar assumption in the human sciences is, to say the least, ill-advised. Mutations within human communities have the habit of becoming highly cointegrated with collective behaviour as people with common interests seek, often through dialogue, to coordinate their subversive acts against conventions that have either been established or are in the process of so being. The presumption that human society's mutation mechanism is 'apolitical' is one of several reasons why the evolutionary take on human history is rather brittle.

¹¹

To their credit, a number of evolutionary game theorists have understood this well and tried to respond analytically. Foster and Young (1990), for instance, acknowledge that politics is what happens when mutations are coordinated into aggregate shocks which test the established conventions. Kandori, Mailath and Rob (1993) examine the impact of rational experimentation in finite and discrete populations. Bergin and Lipman (1996) demonstrate that allowing the mutation probabilities to depend on the current behavioural codes (as opposed to being random and uncorrelated with the present conventions), yields a new type of *Folk Theorem*: i.e. almost *any* conventional behaviour can become disestablished and any alternative may take its place if mutants

coordinate their mutation probabilities appropriately and in response to the current behavioural conventions. This sounds like a celebration of politics as the practice of shaping a society's mutation probabilities and, eventually, of the game. But it also ends all hope that evolutionary theory holds the key to understanding human history. For a theory that explains *all* possible histories as consistent with the evolutionary approach is a theory with very little explanatory power.

Again, I do not wish to underestimate the importance of the evolutionary narrative. It highlights how irreplaceable collective action is in reforming institutions; illustrates how power can be covertly exercised; offers glimpses of how beliefs (particularly moral beliefs) may become endogenous to the conventions we follow; explains how property relations might develop functionally; and so on. What it cannot do is take that crucial, extra step toward genuine historical explanation. But let me be more specific by returning to capitalism's greatest 'innovation' (of taking society from a situation where assets are divided contemporaneously, as in feudalism, to one in which they are distributed inter-temporally). Besides making the whole economic process more productive it made it more reliant on *belief*. The conventions, norms and legal framework of capitalist society had to match the complexity of its technology. Was that complexity qualitatively different to that of preceding social orders in which extractive power was exercised after the crop came in? Did a fundamental shift occur with the transition to capitalism,

viz. the complexity of the socio-economic process? And if so, did the brave new world of capitalist dynamics require new concepts that evolutionary gradualism is ill-equipped to furnish? Can *capitalism's dynamic path be traced through evolutionary models?*

If the answer is affirmative, and all that marked the *Great Transformation* (to borrow Polanyi's, 1945, phrase) was an adaptation of pre-existing norms of distribution, then there is no substance to the claim that history requires a lot more than an evolutionary approach to be laid bare. However, at close inspection of EvGT it seems almost indisputable that history is irreducible to evolutionary dynamics. To begin with, it is ill-equipped to deal even with simpler societies than ours (e.g. feudalism) in which assets are cooperatively manufactured and privately appropriated. At an even more elementary level, it has little to offer the moment the game changes from a simple *hawk-dove*-like interaction over given assets to a fully-fledged *N*-person game of individuals who simultaneously *produce* and *distribute* assets, as well as the social norms that govern these parallel processes. This ought to give us pause: For if farming communities are an explanatory bridge too far for the evolutionary approach, capitalism is even further away from its grasp since the study of systematic extractive power cannot be elucidated by simple evolutionary models which map out the trajectory of behaviour against the background *of a given game*, with *given rules* and *given payoffs*.

At least one thinker thought so long before evolutionary theory made a proper appearance in economics: Marx, who spent much ink describing meticulously the evolution of the *commodity* and of *capital* as analytical categories which cannot be seen as simple, evolutionary adaptations of pre-capitalist phenomena. As commodity exchange became the exclusive means of survival, the commodity-relation replaced human relations. Capital, i.e. the manufactured means of production, '... was not a thing, but a social relation between persons... Property in money, means of subsistence, machinery, and the other means of production, do not yet stamp a man as a capitalist if there be wanting the correlative – the wage worker' (*Capital Vol. 1*, in Marx and Engels, 1979).

The point here is that the whole gamut of capitalist endeavour is based on *particular* social relations. If capital is but a relation-of-production (as opposed to some physical 'thing'), then its value is a matter determined by the network of conventions ruling over

this relation. These conventions, in turn, reflect the *jointly* evolving technologies and relations of production. Steam engines, mechanical looms, and computerised robots are, at once, the secret force behind splendid productive capacity *and* the midwives of our ideology. As technology progresses, it causes raptures in the established behavioural conventions and the associated institutions. Like species that take different forms depending on specific circumstances (e.g. kangaroos that are large and red in the Australian outback but appear small and nippy in the Indonesian forests), a great ecology of capitalist institutions develops around the simple, uniform concepts of commodity production and capitalist accumulation.

The *deep invisibility* and *great variety* of the institutions and social conventions of capitalist production thus play a central role in solidifying both. The resultant

dominant ideology is as uniform and unbending as diverse and multifarious are the various institutions under capitalism. Running through both is a common steal thread: the overarching illusion that observed inequality is not to be explained in terms of the social power of one class over the other but, instead, is the result of different abilities, human capital, work ethic etc. According to this dominant creed, rather than being capitalists or workers, men or women, blacks or whites, *we are all entrepreneurs* (even if some have nothing to sell other than their labour or even their bodies, organs etc.). Indeed, mainstream economics, and by association *game theory*, may be thought of as the highest form of this ideology in the sense that class, gender, race etc. make no sense in the economists' narratives regarding the functioning of the social world.

Our world may have never been so ruthlessly divided along the lines of extractive power between those with and those without access to productive means. It is also more diverse in terms of the coexisting institutions of social distribution than ever. And yet never before has the dominant ideology been so successful at infusing a single idea in most people's minds: the idea that there are no *systematic* and at once *arbitrary* social divisions; that bad inequality is fading fast and that most of the poor are mostly undeserving since talent and application is all the weak need in order to become socially powerful.

12

The question in this paper thus takes its final form: *Can evolutionary theory elucidate the coevolution of (a) astonishing productive technology, (b) deeply entrenched arbitrary discrimination, and (c) the diversity of institutions that help the latter retain the necessary stability by making it invisible?* The answer must be affirmative if social classes, social strata and the pattern of extractive social power in our world can be modelled satisfactorily as by-products of individual interactions guided by the two *statistically independent mechanisms of adaptation and mutation*. I shall now argue that they cannot.

Undoubtedly, EvGT might model, some time in the future, historical change as a feedback mechanism between desires, outcomes, and moral beliefs. Most historians would, nonetheless, require more. The capacity of the human mind critically to reflect on her circumstances and to influence others through dialogue cannot be absent from proper history. Materialist historians would, in addition, demand a special place for the evolution of technologies (and the ecosystem) as a source of non-random mutations closely linked to human inventiveness and political discourse; both sources of collective and individual action that destabilise the prevailing social norms and usher in a variety of brand new institutions. However, for evolutionary analysis to qualify as a source of such insights it must adopt a model of human agency which retains human activity as a positive (creative) force.

13

Can it make room for such an ontology? I think not.

The reason for this negative answer is that the spectre of theoretical *indeterminacy* beckons.

I have already quoted above some of the brightest evolutionary game theorists who seem increasingly pessimistic regarding the evolutionary approach's prospects when it comes to tasks far lesser than the ones discussed in the previous paragraph. As Mailath (1998), a renowned evolutionary game theorist, puts it

[b]oth Refinements [*note* game theory's non-evolutionary, 'conventional', battle against *indeterminacy*] and *evolutionary game theory* were originally motivated by attempts to find the 'right' or 'unique' equilibrium. That hype was not met; and it could not have been met. What has been achieved is a description of the properties of different equilibria.

[*Note added*]

So, even if we were happy to model society as comprising simple automata, instead of creative people, determinate outcomes are elusive. Naturally, the moment we try to complicate the human agent ever so slightly, in order to render her into a historical agent, evolutionary theory is bound to become utterly unhinged.

12.6 On the liberating power of history

One of the implications of the suggested distinction between historical and evolutionary approaches is that the latter cannot furnish a suitable *critique* of evolved institutions. For instance, what is *really* wrong with a world in which the dominant ideology has made most people accept (and even like) the institutions and norms of the prevailing social mode (feudalism, slavery, contemporary capitalism etc.)? An answer along the lines of a moral judgment about the unfairness of the evolved institutions (e.g. of capitalism), based on the observation of inequality and the like, is *not* open to those who are in broad agreement with the evolutionary narrative. For the latter dismisses moral judgements as quasi-illusions functional to the current conventions. The only route available to the critic (who has adopted the analysis so far) is to ground her criticism on something *outside* the evolved belief system.

Marx, for one, focused his indignation on the *inefficiency* of capitalist social relations. His critique of capitalism turns on the argument that it represents a transitory phase of human history; one in which *the social relations* (e.g. the arrangement according to which the set of workers and of owners are, mostly, mutually exclusive) *have not evolved sufficiently to take full advantage of the available technology*. As a result of this mismatch, Marx claims, we live in a society which wastes human resources (in the form of chronic and fluctuating unemployment), devalues humanity (by reducing our relations to commodity fetishism), and requires war in order to maintain some degree of compatibility between (a) what the economy can produce and (b) what consumers have the purchasing power to absorb.

In short, Marx dismisses angrily the notion that capitalism is efficient but unfair, opting instead for the line that it is grossly wasteful of human capabilities, as well as inconsistent with full liberty, because it is *one evolutionary stage behind* the productive capacity of the 'machinery' that it, itself, brought into being. If he is right, it is easy to understand his loathing of both bourgeois *and* proletarian moralities: for they constitute the different sides of the same proverbial coin which

prevents humanity from achieving its potential. Of course none of this requires a slide toward either moral relativism or socio-biological naturalism.

Values matter to humans because of our capacity: (a) to cast a critical gaze on what we do, and (b) to subvert the rules that 'ought' to govern our behaviour, not merely by means of *random experiments* with alternative 'moralities', or codes of conduct, by also by means of *critical reflection, dialogue* and the collective acts that result from these. The point of the rejection of all moralisms is that they circumscribe our capacity to understand the world and, thus, to improve on it. Only, such improvement is made impossible if, along with the moralistic bathwater, we throw away all values furnished by History.

Regardless of whether one agrees with Marx's overall critique, the question which he posed implicitly, regarding the contest between evolutionary and historical approaches, and the possibility of an ethical critique of our institutions, is important and remains unanswered:

How can we *criticise* our social order (slavery, feudalism, capitalism etc.) without resorting to the normative views that have been foisted upon us by that very social order (and the preceding ones whose moral codes remain somewhere deep in our conscience; just as our appendix is a relic of some primitive incarnation of our species)?

Evolutionary theory has, despite its undeniable merits and overall oeuvre, some natural limitations beyond which it cannot reach: regarding critical reasoning, moral judgements, and normative beliefs, all it has to say is that they are illusions functional to our *given* interests, which we pursue within *given* games, and under *given* rules. Although there is a great kernel of truth in this, history cannot tolerate so many *givens*. By moving beyond them, historical approaches inspire hope of liberation from our illusions without, however, pushing us into the sinister embrace of moral relativism. The Study of History (and perhaps of Art and Music) delivers us from artefacts of our own creation which (once milked for all they are worth) we must transcend. An example of this comes in the form of a sentence that we ought to put to evolutionary game theorists:

How is it that you can explain moral beliefs in materialist terms, but you avoid a materialist explanation of beliefs about what we consider to be our in own interest? If we are capable of having illusions about the former (as you admit), surely we can have some about the latter! If morals are socially manufactured, then so is self-interest. If institutions play a role in what we consider our self interest to be, then people populate institutions as much as institutions populate... people.

12.7 Epilogue

Institutions distribute social power. Neither the former nor the latter make sense in a historical vacuum. Without a decent account of how they spring out of the social conventions ruling over our practices and our beliefs, institutions will remain opaque and social power as invisible as it is all-encompassing.

[Section 12.2](#)

examined some theoretical and experimental insights from evolutionary game theory *vis-à-vis* the establishment of social conventions and their concomitant ethical beliefs in communities where accumulation is of a primitive type.

[Section 12.3](#)

argued that food production laid the foundation for the simultaneous evolution of conventions determining property rights over land and extractive power over the collectively produced surplus.

[Section 12.4](#)

focused on the crucial difference brought on by capitalism: the reversal of who claims the residual and who gets their share first (the residual claimants or the rest?). It showed that the variety of institutions of capitalism is fully consistent with the evolutionary approach and argued that, despite this great variety, capitalism engenders a uniform, single ideology regarding the illusory causality between privilege and 'worth', or 'virtue.'

[Section 12.5](#)

then argued that history (especially that of capitalist societies) is irreducible to evolution and that the evolutionary approach is insufficiently evolutionary. Finally,

[Section 12.6](#)

added the speculative claim that the historical approach possesses a liberating capacity that the evolutionary approach lacks.

One of the insights that came to the surface while scrutinising evolutionary game

theory's capacity to illuminate historical change relates to the very notion of social power. It was the thought that historical developments of great note (e.g. the institutions of food production and capitalist relations of production) boosted handsomely the degree of extractive power exercised by elite on non-elite groups while, at the same time, shrouding social power in a veil of obfuscation. And yet history moves on. How does that happen? Is there anything other than technological and ecological change that destabilises established conventions of social power, thus giving the wheel of history another twirl? There is, I argued. It is the tendency of humans to reflect critically on their actions and to subvert collectively the norms that, supposedly, 'ought' to be ruling their behaviour; a tendency which is at least as natural (even if less frequent) as the tendency to conform. The effect of this tendency is to keep conventions of social power constantly on their toes, ready to subvert them the moment some technological or other development has upset their evolutionary fitness. No genuinely historical approach can afford to leave out of its ambit a model of humans as creative agents capable of both individual contemplation and collective subversion.

Seen from this perspective, neoclassical economics is a deeply regressive project. For more than a century it tried to keep history out of its analyses, portraying capitalism as a timeless realm of pure exchanges. Then, when advances in game theory combined with mathematical biology to allow for evolutionary processes to be admitted into mainstream economics, the profession made sure that history would continue to be denied a foothold. How? Simply by assuming, via a fresh variant of neoclassicism's third meta-axiom, that mutations cannot be patterned; that they must be identically, independently and randomly distributed!

Whereas biologists have good cause to make this assumption (as the adaptation process of genes and memes alike is, in the animal republic, distinctly separable from the process that throws out mutations, randomly and unpredictably), a similar

assumption in the social sciences is tantamount to an embargo on politics. For what else is politics if it isn't about patterned mutations? While acts of subversion can be individualistic, society changes fundamentally only when people coordinate their subversive acts in a manner that destabilises established norms. This is precisely what made Spartacus, trades unions, Martin Luther King Jr, the feminists etc. significant historical figures with a lasting legacy upon the present.

VERDICT: There is a fascinating connection here between neoclassicists' assumption that mutations *must* be uncorrelated with their assumption (recall

[Chapters 3](#)

and

[4](#)

) that bluffs can never really work because the probability of stepping out of the neoclassical behavioural equilibrium is commonly known. The connection is none other than neoclassicism's third meta-axiom which imposes equilibrium on fundamentally and profoundly disequilibrium phenomena for the cynical purpose of 'closing' the model. Just as in the centipede game of

[Chapter 4](#)

, so too here the assumption that deviations from equilibrium (whether bluffs in bargaining or mutations in some social evolutionary process) must *always* and *necessarily* be empty of meaning (i.e. random, unpatterned, free of political significance, vacuous noise as opposed to meaningful signals) is instrumental to the needs of finding a determinate solution to the neoclassical model. The fact that there is no logical reason to make this assumption is neither here nor there for the neoclassicist. The further fact that this is an assumption that debases history, eliminates genuine freedom and belittles human reason is, for the neoclassicist, acceptable collateral damage.

Appendix 12.1: The evolution of mixed strategies in the hawk–dove game when the population is homogeneous

Suppose that in the *hawk–dove* game the population is homogeneous (i.e. all players are identical) and initially programmed to play strategy d in the game. In each interaction all players retreat (i.e. play dovish strategy d) and each receives payoff 1 every time. Suppose now that, for some unspecified reason, one player switches to strategy h . This ‘switch’ could be a ‘tremble’ or ‘error’ or (in the language of biologists) a ‘mutation’ (that is, the rare birth of a ‘hawk’ to a dovish parent). Alternatively we may think of it as an experiment performed by an inquisitive ‘dove’. Whatever the reason, a lone player selecting strategy h in a population of ‘doves’ will collect payoff 2 in each interaction (with her opponent collecting 0). If this relative ‘success’ translates into relatively more offspring to our ‘mutant hawk’, or if other doves mimic the mutant’s relatively successful strategy and turn into hawks, the proportion p of h -playing agents in the population will grow. In this sense, evolutionary biologists tell us, an homogeneous population of d -players is susceptible to an *invasion* of h -playing mutants. Outcome dd is, consequently, evolutionarily *unstable*.

The same applies to outcome hh . For if all players are initially programmed to play h , the cumulative payoffs of a d -playing ‘mutant’ will be higher than the norm. Thus, generalised h -playing (d -playing) cannot survive evolutionary pressure in an homogeneous population as proportion p falls (rises) following

the birth of a mutant dove (hawk). In short, if p is too high, evolution will force it (via mutations and copying of relatively successful strategies) to fall while if p is too low it will tend to rise. When will it stabilise? The answer is: When p equals exactly $1/3$, which coincides with the *Nash equilibrium in mixed strategies* (NEMS). To see this, check when $p < 1/3$ the expected payoffs from h [i.e. $-2p + 2p$] exceed those from d [i.e. $0p + p$]. When this happens, p grows as the proportion of hawks increases. And vice versa. Thus, the proportion of hawks stabilises when p is neither larger nor less than $1/3$; i.e. when $p = 1/3$. Once at that evolutionary equilibrium, in each round there is a probability of $1/3$ that each person will play h .

Appendix 12.2: How an arbitrary feature makes arbitrary discrimination evolutionarily inevitable

Suppose that the population is heterogeneous; there are two types of player distinguished by some arbitrary feature: one group consists of ‘blue’ players and the other of ‘red’ players. Suddenly, players can condition their behaviour on their opponent’s arbitrary colour. The potential thus exists for heterogeneous behavioural codes. To the protests of conventional game theory, that there is no rational motive to condition one’s behaviour to one’s opponent’s meaningless colour, EvGT retorts that successful strategies need have no good reason behind them other than their... relative success. The question for EvGT is: Will conditional strategies (that is, strategies which instruct players to play differently against opponents of different colour) gain an evolutionary upper hand over unconditional ones? Starting from the equilibrium described in

[Appendix 12.1](#)

, where each player plays h with probability $p = 1/3$, we note that since there are now two types of player, we need to mark separately the probability that a ‘blue’ will play h from the probability that a ‘red’ will play h . Let’s call these, respectively ρ and β and assume that, at the outset: $\rho = \beta = 1/3$. Any mutation that alters either ρ or β ever so slightly will push the population’s behaviour into one of two situations: Either the ‘reds’ will be more aggressive or the blues. Once some mutation causes ‘blue’ players to become slightly more hawkish than ‘red’ players, an evolutionary bandwagon will begin which will lead to an evolutionary equilibrium (EE) such that in meetings between

'blue' and 'red' players, either all 'blue' players act aggressively (i.e. play *h*) toward and 'red' players, who acquiesce (i.e. play *d*), or the opposite. See Hargreaves-Heap and Varoufakis (2004, Chapter 6).

Notes

- 1 See Varoufakis, Halevi and Theocarakis (2011).
- 2 This chapter is based on Varoufakis (2008).
- 3 Malthus was concerned that human population grew geometrically while food production could only grow arithmetically. If so, a struggle for existence would occur as increasing numbers of people would have to starve. Darwin (1860) was clearly impressed by this, in his own words: 'In the next chapter the Struggle for Existence amongst all organic beings throughout the world, which inevitably follows from the high geometrical ratio of their increase, will be treated of. This is the doctrine of Malthus applied to the whole animal and vegetable kingdoms' (pp. 4–5).
- 4 There is, indeed, something endearing about the excitement which EvGT has brought to a profession which abstained for so long from any evolutionary or historical investigation courtesy of its self-imposed analytical constraints. EvGT has had something of a liberating effect on the economists' imagination and this must be welcomed.
- 5 The idea here being that an injury reduces one's chance to reproduce.
- 6 In the sense that they were making choices in an environment extracted from their social setting; one in which no value was produced and interaction was defined fully by serial contests over \$2 'pies'.
- 7 In these fresh experiments, the 32 rounds of the *hawk–dove* game (where players had been assigned, randomly, the blue or the red label, as in Chapter 11) were followed by two rounds of the so-called *ultimatum game*. In the latter one player is asked to offer the second player a division of \$10 on the condition that, if the second player rejects the offer, no one gets anything. Noting that the *Ultimatum Game* is the simplest type of bargaining (or cooperative) game, it is interesting to report that the players whose colour in the earlier part of the game (while *hawk–dove* was being played) had emerged as dominant were offering far less to players of the 'other' colour and vice versa. It seems, therefore, that the iniquitous institution that evolved in response to the primitive accumulation game (the *hawk–dove* interaction) spread by analogy to the *ultimatum game* in a manner not dissimilar to that described theoretically by Sugden (1986, 1989).
- 8 As I shall be arguing below, capitalism propelled extractive power to its apotheosis courtesy of the extreme asymmetries it introduced between the outside options of (a) owners and (b) non-owners of productive means.
- 9 A choice between hunger and selling one's labour to the highest bidder is not an easy one. But it is still a choice compared to the 'choice' between being stabbed and handing over a part of one's output to an aristocratic residual claimant.
- 10 In particular, one would not expect a convention which generated relative losers and which confined them to the interactive margins (that is, placed them in a position where they were less likely to interact with their fellow adherents) to last long. Or to put the last point even more simply, where conventions create clear winners and losers, two conventions are more likely to coexist when communication between followers of different conventions is confined to the winners of both.
- 11 I wish to thank Geoff Hodgson for assistance in the elucidation of this point.
- 12 For a discussion of 'good' versus 'bad' inequality see Varoufakis (2002).
- 13 Marx habitually poured scorn on those (e.g. Spinoza and Feuerbach) who transplanted models from the natural sciences to the social sciences with little or no modification to allow for the fact that human beings are very different to atoms, planets and molecules. We mention this because at the heart of EvGT lies a simple Darwinian mechanism (witness that there is no analytical difference between the models in the biology of John Maynard Smith and the models in this chapter). Of course Marx himself has been accused of mechanism and, indeed, in the modern (primarily Anglo-Saxon) social theory literature he is taken to be an exemplar of 19th century mechanism. Nevertheless he would deny this, pointing to the dialectical method he borrowed from Hegel and which (he would claim) allowed him to have a scientific, yet non-mechanistic, outlook.
- 14 I am hereby referring to the fact that evolutionary theory gets bogged down in a plethora of evolutionary equilibria. none of which more likely to emerge than the rest.

13 Conclusion

Dealing with indeterminacy on the stage of social life

13.1 The social theorist as storyteller

Now that all is said and done, and this book on my 'personal encounter' with the discursive power economics builds on its gross theoretical failure is complete, perhaps the reader will allow me to conclude on a light-hearted, yet deadly serious, note.

Social theorists, whether we like it or not, are storytellers. We tell meta-stories the purpose of which is to help us understand a social world that is constantly under construction around us, and within which we are both active contributors and incessant interpreters. Incapable of a genuine Archimedean perspective, from which to judge simultaneously both our world and our account of it, we resort to analysing and re-analysing the sort of stories we tell ourselves about ourselves.

This book has focused on one particular species of meta-story: that which dominates economic textbooks and discourse from the high school curriculum all the way to finance ministries and the boards of the too-big-to-fail financial behemoths. Having already subjected to critical scrutiny the neoclassical meta-story, and the dance of the meta-axioms which keeps it powerful and alive, I shall now end this book with an odd question: Given that drama is the ultimate, and most instructive, form of story-telling, what type of play or novel could a neoclassicist come up with without violating her neoclassical strictures? What would indeterminacy's role be on her stage, or in the pages of her novel?

13.2 The neoclassical economist as playwright

1

It would not be remiss to imagine that the neoclassical economist tries to be for social theory what the playwright is to theatre: the creator of the plot, the designer of its every twist and turn, the author of a morality tale. At first, she introduces us to the cast of characters, their whims and preferences, their constraints and social location. Next, she sets the scene by developing the intricate web of interdependent decisions that forms their milieu – the grand, usually competitive, 'game' they are engaged in.

To pack dramatic punch, this grand game must feature multiple equilibria. Let me explain why: George Bernard Shaw, in the preface to one of his exquisite

plays, wrote: 'No conflict, no drama!' Hear, hear! To fashion genuine drama, our neoclassical playwright will undoubtedly feel compelled to weave an authentic conflict into its fabric. For this to be so, the 'game' that she places her characters in, must, so as to stay faithful to neoclassicism's tenets, feature more than one equilibrium outcome. Imagine for a moment that it does not, and that each of her characters possesses a dominant strategy. If so, they will be facing no real dilemmas. Their dominant strategy may well instruct them to slaughter and to pillage, to wreak havoc and turn other people's lives upside down. But the result will be crude, brutal theatre. Conflict contributes to a decent play only by allowing the audience to identify with even the worst villain. Interesting, mesmerising conflict involves a degree of inner turmoil the prerequisite of which is some form of... indeterminacy.

The neoclassicist's conundrum here is, naturally, that indeterminacy is her sworn enemy. And yet, in attempting to write her captivating 'play', she realises that she cannot dispense with at least *some* indeterminacy. It is for this reason that multiple equilibria are, as I stated above, *sine qua non* for neoclassical theatre. The question now becomes: What constraints does a reliance on multiple equilibria impose upon the neoclassical playwright? We know that, by the very nature of neoclassical method, once the multiple equilibria are in place, the playwright's characters have no alternative other than to resolve them by means of some form of randomisation. For if some other non-

random mechanism were invoked, which rationalises in front of an audience the selection of one out of multiple equilibria, the whole point of having created a play with multiple equilibria is defeated.

Indeed, if the play's leading character can reliably select a path among many competing ones *without resorting to randomisation*, it must be the case (at least in the neoclassical mind-set) that this path is, by definition, optimal and thus consistent not with multiple but with a unique equilibrium. But if this were so, the play is constitutionally dull, featuring characters whose path is predetermined (by the unique equilibrium available to them) and whose choices are, consequently, unrevealing of anything that may cause a ripple of excitement through the audience, or help them re-assess their own humanity.

Creative neoclassical playwrights may attempt to escape this conundrum by allowing their characters occasionally, and in the heat of the moment, to depart from instrumentally rational choices. In the parlance of game theory, the neoclassical playwright may introduce some 'trembles', or noise, in the characters' behavioural pattern, hoping that this will complicate the plot sufficiently for the audience to find the proceedings absorbing. The question is: How absorbing would such a play really be?

I harbour no doubt that many Hollywood flicks are written in more or less this (neoclassical) fashion. In the tradition of John Wayne and Arnold Schwarzenegger movies, the screenplay is mostly predictable and the characters' actions predetermined, save perhaps for a little randomisation reflecting the possibility that those in the weaker position err randomly into thinking that they stand a better chance than they really do (before being mowed down by the 'hero'). However, such examples of 'drama' would not satisfy a sophisticated neoclassicist playwright or

screenwriter who, outside her professional world, knows the difference between a good play (written for the screen or stage) and an impostor. The reason why offers a useful platform from which to reconsider neoclassical economics in general and its highest form, game theory, in particular.

At first, there is the problem with the use of multiple equilibria as a device for raising the dramatic tempo. How does the playwright resolve them? The equilibrium solution, demanded by neoclassicism's third meta-axiom, is to have a unique mixed strategy equilibrium cutting the Gordian knot of indeterminacy by some optimal randomisation rule. Or that protagonists reach a settlement reflecting a Nash bargaining solution that equalises ratios of their utility functions' first- and second-order derivatives. It would be *as if* Shakespeare decided on whether Macbeth would commit murder by tossing a suitably biased coin, or as if Sophocles had Antigone and Creon resolve their 'disagreement' by maximising the product of their utilities, and then writing the play according to the outcome. If our playwright rejects the mixed strategy Nash equilibrium solution (as I did in various chapters, beginning with

Chapter 3

), then the equilibrium story will either be devoid of tension (as the playwright will fall back on dominant strategies) or it will require external input for its completion (e.g. some *deus ex machina* that introduces reasons, for the outcome, that are external to the 'game' the playwright set up). In view of our playwright's need for tension but also her penchant, as a committed neoclassicist, for endogenous 'closure', she is left with no good option.

One escape route, suggested by the last paragraph, might be to consider the neoclassical method as the play's grammar or background logic while the playwright's imagination is permitted to look to some external source of inspiration for the actual course that the characters will trace out within that 'grammar' or 'landscape'. Under this partial retention of equilibrium theory, the playwright is at liberty to deploy mixed strategies, Nash bargaining solutions, general equilibria, conventions that evolve in

ways Evolutionary Game Theory can understand (recall the last chapter). However, the secret of the play is not to be found in any of these theoretical constructs. Although a play is possible that relies entirely on them, I for one would not enjoy it much. And I do not think that our neoclassicist playwright would either.

13.3 Rational deviations from equilibrium as the prerequisite for good theatre

So, the identification of taxing dilemmas with multiple equilibria does not have to lead to boring plays, as long as the playwright first sets up the dramatic structure and then augments the equilibrium path that the dramatic structure generates with exogenously determined events. However, I fear that ushering in resolutions which are independent of that structure (as they would have to be, since the equilibrium foundations can neither depend on the resolution nor spawn a resolution) would not work at all well. This kind of play would require that the dramatic tension be structurally independent of its culmination. For those of us who think that a *deus ex*

machina resolution is detrimental to good theatre, the only decent alternative is to abandon the synonymy between hard choices and multiple equilibria, just as I abandoned the identification of the equilibrium strategy in the centipede game of

Chapter 4

with the uniquely rational course of action.

Indeed, the discovery of rational non-equilibrium strategic choices in various chapters of this book (i.e. the possibility of perfectly reasonable patterned bluffs, mutations and deviant acts) introduces us to a different class of dramatic devices from which the playwright can build satisfyingly three-dimensional characters. The very possibility that Antigone can defy her dominant strategy rationally is what gave the famous play its genuine dramatic texture, focusing as it does the audience's attention on not just Antigone's dilemma but, also, on its own inconsistent views regarding what is rational, ethical and, in the final analysis, 'right'. On the edge of their seats, the audience suddenly expect the unexpected at every moment since 'uniqueness' no longer guarantees determinism. They anticipate dramatic choices not only when the alternatives are equally inviting for the characters but even when they are not!

This is precisely the essence and beauty of deviant, yet fully rational, strategies; of the strategies that neoclassicism attempts to eradicate by means of its third meta-axiom. By undermining the characters' perception of what rational people do, subversive behaviour may or may not reward those who adopt it with superior outcomes. On this account, Aeschylus raised Prometheus from the obscurity and banality of the Hesiodic tale by empowering him to experiment with a subversive, a deviant, an 'out-of-equilibrium' strategy. Of course, other more cynical interpretations are always possible. Prometheus could simply be suffering from imperfect information, in which case he would not have repeated his gift to humanity had he been given a second chance. Alternatively, he could have enjoyed martyrdom and accepted pain as a reasonable price for it. However, both these interpretations cheapen the character created by Aeschylus. No tragedy is worth its salt if it is based on the exploits of a hero who simply miscalculated, or who unexpectedly learnt how to derive masochistic pleasure when things turned differently to his original plan.

The above musings lead gradually to the conclusion that high theatre is impossible without radical indeterminacy, i.e. without neoclassicism's nemesis which economists habitually, mostly unwittingly, try to put away by means of their dance of the meta-axioms. But even if we agree that indeterminacy is a prerequisite for a good play, it is certainly insufficient. Indeterminacy prevailed in

Chapters 2

to

, when the agents' motivation was taken for granted. However, unless their praxes infect their motivation and influence their beliefs, there is no real theatre to speak of.

Chapters 5

to

took account of this interdependence which, at once, builds on indeterminacy and takes it onto a higher plane of complexity. That escalation and decent into ever more radical indeterminacy was, at least from a self-respecting playwright's perspective, apt and timely.

As every schoolchild knows, Macbeth adds crime to crime, as a result of successive choices that he 'fell' into when 'murder' was one of several options. He then emerges defeated while simultaneously victorious. When he achieves clarity

toward the play's end, he tells us that he wished his 'beliefs' were expunged:

Canst thou not minister to a mind diseas'd,
Pluck from the memory a rooted sorrow,
Raze out the written troubles of the brain,
And with some sweet oblivious antidote
Cleanse the stuff'd bosom of that perilous stuff
Which weighs upon the heart?

(Macbeth, V, iii, 40–4)

But at the same time he re-discovers his dignity when forced to choose between death and humiliation:

Lay on Macduff,
And damn'd be him that first cries, 'Hold, enough!'

(Macbeth, V, vii, 62–3)

The neoclassical interpretation is that Macbeth became involved in conflict due to imperfect information as to his opponents' preferences and strengths; that he made his choices rationally given his priors of belief; that his choices had been hard because they belonged to sets of equilibrium strategies containing more than one element; that, perhaps, there was an element of the irrational in his deeds that took the form of some random 'tremble'; some deviation from the optimal strategy which packs no significance in itself; and so on. All these suppositions are contentious because they wilfully ignore the possibility that a rational Macbeth could have chosen from the non-equilibrium strategy set as well (as I have contended in this book). None, however, is as contentious, and downright absurd, as the assertion that Macbeth's choices left his personality unaffected, or that they affected him in a predetermined (even if stochastically so) manner. Once choice and action contaminate motivation, the latter cannot determine the former. Indeterminacy is, therefore, irrepressible not just as the stuff of good theatre but also as a prerequisite for human development.

13.4 Persons and roles, impersonators and actors

The importance of praxis for deciding what kind of theatre (and social theory) we want becomes apparent when the repercussions of the acceptance of the neoclassical equilibrium story are considered. If tensions and dilemmas are the result of multiple equilibria, their resolution cannot be seen as a contributor to the character of agents. Macbeth's tragedy is not so much about his fate and final destruction (i.e. it is not about the outcome *per se*) as about the transformation of his self through praxes – a transformation that the neoclassical story cannot even begin to conceive. Even if we approach the matter from an evolutionary game theory perspective (recall the last chapter), the neoclassical insistence that character variety is spearheaded by statistically uncorrelated 'mutations', or replication errors, amounts to the same thing: praxes are not causally linked to the evolution of character.

If we are to make room for this two-way process, from self to action and from action to self (the process that I call praxis), theatre regains its potency and meaning, although the neoclassical project sinks without a trace in a sea of radical indeterminacy. But what kind of theatre do we end up with? One possibility is a relativist open-ended script, in which the playwright lets the actors improvise and bring on the stage perspectives from their own lives. Together with neoclassical narratives concerning the outcome, such 'libertine' improvisation brings an end to *any* type of theory about how the play will end. Suddenly we are no longer in the realm of theatre but in something much closer to a multi-player video game in which the game's author offers the environment but it is the players that collectively script the outcomes.

The input from the social world, the world of actors, or gamers, who have a life outside the 'game', or play, decides the outcome and there is no sense in considering the script as anything more than a basic skeleton which cannot be read independently and across different cultural milieus as transcendental (unlike for instance a play authored by Shakespeare or Sophocles). It is this post-modern relativism that neoclassicists of renown (particularly game theorists such as the formidable John Harsanyi) have sought to keep at bay by excluding from games anything that was not in the players' utility payoff functions (that is, in the 'script'). But he may have gone too far, killing off the very possibility of good theatre, of wholesome social theory, or even of a believable model of men and women. Is there an alternative to the choice between bad theatre and multi-player video games?

I think so. The claim buried deeply inside this book is that creativity and interest (in theatre as well as in the social sciences) can be restored without going to one of these extremes. There is, I submit, no need either to succumb to deterministic scripts or to go all the way to the other extreme, yielding to relativist, arbitrary, actor-imported narratives. The theory of deviant-*cum*-rational behaviour that unfolded through various chapters of this book opens up the possibility of having a complete, riveting script without determinism. The key is the admittance to the plot of twists that *defy* unique equilibria and in so doing capture the imagination. When one is faced with multiple equilibria, any odd choice will do; by observing the outcome you really learn nothing about the chooser's character. It is only when the character reasonably and purposely deviates from some unique equilibrium path (akin to Prometheus' theft of fire on behalf of humanity) that the writer is telling us something momentous about both the character and the human condition.

The readiness to infuse rationality with irrationality in the defence of deviance and rebelliousness brings to mind another dimension. In a 'neoclassical' theatre production the actor sheds her own personality, steps into the predetermined role, and struggles to *impersonate* some character whose entire presence flows from the script. The playwright chooses how to weave the plot and the actor attempts to be

the person who actually does the weaving. However, in high theatre, rather than impersonating characters, actors *personify* them.

My preference for scripts which depict and rationalise non-equilibrium choices is well suited to complementing this distinction; to actors that personify rather than merely impersonating. This way, actors can bring on stage their own interpretation of the inner conflict that characterises the choice of some deviant strategy with no need for an open-ended, relativist, multi-player video game-like, script. Since the coexistence of Reason and Unreason in the *same* behavioural pattern (recall

Chapter 6

) is inherently confusing, the actor can draw from her own past whatever is necessary to convey the intensity of such choices. Neoclassical theory cannot (because it *will* not) account for that intensity, thus devaluing, quite inadvertently, the actor's contribution to the play.

Does any of this matter? I think it does, and that Macbeth offers a good example. For he demonstrates that, often, it is obligatory that in order to understand conflict in the social world around us, we must look into a conflicted heart overseen by a bright mind. If we are to do so without abandoning the realm of rational analysis, the only option is to enrich our perception of it. Rejecting neoclassical equilibrium analysis is an excellent start along this path.

13.5 Toward a manifesto for an indeterminate economics for the post-2008 era

Theatre offers a vivid illustration of the main consequence of divorcing Reason from Knowledge, and leaving Knowledge in the domain of Natural Science. Neoclassical plays, based on this separation, would amount to a succession of interconnected scenes, the suggestive power of which is exhausted when one finds out what happens in the final act. Similarly, neoclassical art, if it can ever be imagined, would present itself as an escape into the realm of the refreshingly irrational; an escape that is harmless when confined to a Museum of Modern Art, or to a concert hall, but which is thought of as supremely dangerous if allowed to infect 'serious' decision making.

The Crash of 2008 ought to have exposed, once and for all, the sad truth that the 'serious' decision-making of 'very serious people' was never founded on rationality. That, instead, it was predicated upon a type of putrid expediency which received substantial ideological support from the pseudo-scientific discourses of neoclassical economics. The rejection of this toxic theoretical approach, that is neoclassical economics, allows us, once again, to see art and theatre in a different light. They are no longer benign and refreshing celebrations of the irrational but, rather, studies in the art of good choices; they share the same method and quality with helpful, insightful social theory. Nonetheless, although art reinforces the need for a departure from neoclassicism's model of men and women, which it has so obediently been modelling on nineteenth century classical mechanics, it is to radical indeterminacy that we must yield, and which we must embrace with creativity and expectation.

Throughout this book, I have been narrating my personal encounter with the irrepressible indeterminacy that neoclassical economics reliably throws up every time it tries to offer a slightly less unsophisticated view of capitalism. On each and every occasion, as evidenced in the preceding chapters, neoclassical economics responded to the resulting indeterminacy with vicious abandon, either by ignoring it or by trying to bleach it out of its models by means of its third meta-axiom. Either way, economics extended its social power over the rest of society through a thinly disguised form of intellectual fraud, converting itself in the process into a kind of theology with equations.

The question now becomes: If I am right, what is the way forward? Put simply, my final missive to interested readers is that we need to embrace indeterminacy *without losing our affection for theory*. In a recent book jointly authored with Joseph Halevi and Nicholas Theocarakis,

²

we included (see

[Chapter 10](#)

) a section entitled 'The primacy of radical indeterminacy'. I can think of no better way of ending this chapter than by reproducing that argument here. Our critique, in that book, extended well beyond neoclassical economics. Indeed, we argued that there is a general pattern afflicting all schools of economic thought, some more than others of course: When economists try to squeeze consistency out of their models, the result is always and reliably grand failure, we argued. While such failure can leave economists' job prospects unaffected (or even, sometimes, enhanced), eventually it deprives the theory of persuasive power.

For example, David Ricardo's insistence of squaring his value theory with a theory of

growth led to Malthus' devastating critique; Marx's desperate attempt to close his model led to the *transformation problem* and the contorted logic required for its resolution; the neoclassicists' insistence of explaining all prices and quantities by means of the equi-marginal principle forced them, eventually, to stick to Robinson Crusoe-like economies, etc. The trouble with these failures was that, once their logical incoherence became apparent, and the political order no longer had uses for them, they led the following generations of economists to drop them wholesale, together with the important insights contained within. And if this has not happened just yet with neoclassicism, because of its continuing political utility, eventually it will.

The tragedy of economics is, therefore, that the economists' unwillingness to acknowledge indeterminacy, and their determination to push it under the proverbial carpet, leads to a whole sequence of *lost truths*; insights about capitalism that were, once, better known by previous generations of economists. So, whereas even right-wing economists at the turn of the twentieth century benefitted significantly from Marx's thought (e.g. Joseph Schumpeter who has acknowledged his gratitude to Marx for the development of his idea of 'creative destruction'), today's crop has no access to such truths. Oblivious to the lessons learnt by previous generations of political economists, they march straight into their own theoretical Waterloos.

What should we do, in view of this repetitive process of theoretical failure, caused by the resistance to acknowledging indeterminacy? My answer and recommendation (which has been put forward in a number of books),

3

is truly simple:

Adopt Sisyphus's optimal strategy! That is, stop pushing the rock up the hill. Just embrace indeterminacy and stop employing techniques of exponentially increasing complexity in order to elevate it onto a higher plane, without ever eradicating it. The impossibility of the task of weeding indeterminacy out should not give us extra energy to tackle it but ought to grant us pause to think of that which constitutes our real task: To explain the social world, capitalism in particular, as an inherently indeterminate system.

But this means a complete disengagement from the impossible project of discovering the truth about capitalism within some determinate abstraction; within some 'closed' model. In methodological terms, this is equivalent to abandoning rigid meta-axioms even if the price we have to pay is *radical indeterminacy*. Would the latter constitute a serious defeat, as neoclassicists are convinced it does? As this book has shown, it constitutes no such thing: For even when we impose the most stringent of meta-axioms, *radical indeterminacy* cannot be defeated. Why then pay the price exacted by the meta-axioms (that is, total historical blindness and a sequence of serious violations of logic) when, in truth, they do not even deliver us from indeterminacy? Clearly, no good reason presents itself.

Having said that, is there a precedent of embracing radical indeterminacy in economics? Yes there is but, unfortunately, the said embracing has always been incomplete.

Table 13.1

offers a summary of how different schools of economic thinking have stood in relation to indeterminacy. The first three columns correspond to neoclassicism's three meta-axioms, as discussed in

Chapter 1

and beyond. The last two (entitled 'spontaneous order' and 'reducibility of human action') refer, respectively, to (a) the meta-axiom which has it that decentralised, unregulated, market-driven behaviour achieves social welfare-enhancing order, and (b) the meta-axiom that men and women are analytically equivalent to machines; to a mathematical mapping of outcomes to some index of preference satisfaction, or, at best, to the algorithms running our magnificent computers.

As for the table's rows, each corresponds to a major school of economic thought: the first one represents the Adam Smith and David Ricardo classical economics tradition (which is also espoused by more modern thinkers like Pierro), the second row is occupied solely by Karl Marx, the third is dedicated to early nineteenth century marginalists (e.g. Jevons), the fourth row is populated by the marginalists that turned neoclassical (through the adoption of the third meta-axiom, as discussed in

Chapter 1

), the fifth row refers to marginalists of an Austrian tradition, and the sixth row accommodates John Maynard Keynes. The reader will also notice that there is a seventh row, left unbranded: it is the one to which I feel that I belong – possibly a minority of one...

As this book was devoted to a long tirade against neoclassicism, perhaps it is best to begin with the fourth row that corresponds to that school. All boxes are ticked, except for the spontaneous order one. Let's see why: The first three columns correspond to neoclassicism's three meta-axioms. It is therefore evident that they are 'checked' by the neoclassical tradition. Of the remaining two

columns, while the last is also checked (since the neoclassicists are renowned 'reductionists') the penultimate one, corresponding to 'spontaneous order' is 'crossed'. Indeed, neoclassicists do not automatically assume that the best of all possible worlds will automatically emerge if all decisions are decentralised (i.e. they do believe that asymmetrical information and monopoly power can be sustainably detrimental to social welfare).

The difference between the neoclassicists and their forefathers, the nineteenth century marginalists (the third row), is that the marginalists did not go as far as to espouse the neoclassical penchant for imposing equilibrium axiomatically (as opposed to explaining convergence toward equilibrium, e.g. Cournot or Marshall). And since economists of the Austrian persuasion and John Maynard Keynes also locate their roots in that form of marginalism, while steadfastly refusing to *assume* equilibrium a priori, they too 'get' ticks in the first two rows, but not in the third.

In contrast, David Ricardo, the neo-Ricardian Pierro Sraffa and Karl Marx, like all classical economists, make no assumptions about individual *agency*, and thus get crosses in the relevant boxes (see the first two rows under **D** and **S**). Of course, courtesy of their imposed assumption that inter-temporal equilibrium prevails in the macro-economy (as the rate of profit tends to equalise across the different sectors; and supply equals demand for all produced commodities), they too get ticks in the third column (**E**).

To sum up, the first two columns (meta-axioms **D** and **S**, which were fully explained in

Chapter 1

) typify an individualist approach to *agency*. In such accounts, *structure* is to be explained by an *agency* located in individual action that is instrumental and comes prior to *structure*. The next two columns concern the manner in which the theorist comes to firm conclusions about *regularity*, without which no firm predictions can be made (and regardless of whether the agency boxes are ticked or not).

There are two ways in which we can extract *regularity* from a theoretical model: The most common one is through the strong version of neoclassicism's third meta-axiom (**E**); that is, by assuming equilibrium not only exists but, additionally, that it is the only state of the economy worth studying. Interestingly, both neoclassicists and classical economists, including Marx, took that step.

Meta-axiom **E** is, arguably, so strong and logically unwarranted, that a number of marginalists refused to espouse it. The first to refuse **E** was Cournot (1838), who even sounded a warning to the effect that humanity might embark upon a lethal path if **E** is

endorsed, together with **D** and **S**. Beyond Cournot, the Austrian school turned **E** down, perhaps because of the fact that their point of origin was a critique of Marx's espousal of **E**. Nevertheless, since they were just as politically driven as Marx (even though they were trying to make precisely the opposite point to his), they too craved *regularity*. For without *regularity*, no theory has firm predictions. And without firm predictions, how can a political economist advocate particular policies?

For this reason, the Austrian School came up with an interesting alternative to **E**: the idea of a *spontaneous order* that is 'as good as it gets'. They begin their narrative by endorsing the first two meta-axioms (which define human ontology

and the way we must conduct economic 'science') but they reject the notion that some equilibrium will result. For if it could, human Reason might be able to work out what that equilibrium would be and, then, socialism might be justifiable (as a system that imposes that very equilibrium).

To render socialism wholly indefensible, they had to argue that equilibrium is neither possible nor desirable. Thus they put forward the hypothesis that, due to the irreducibility of human knowledge to some well defined mathematical function, no central plan and no collective agency (i.e. a state, a municipality, a club) can generate social outcomes. The best humanity can hope for is the social outcome that will emerge *spontaneously* if people and markets are 'left alone'. Thus, the Austrians sought *regularity* in the *spontaneous order* resulting from free intercourse (a meta-axiom we label **O** in

Table 13.1

) between persons (who are to be theorised on the basis of the first two meta-axioms, **D** and **S**).

The Austrians were not the only ones to reject **E** while embarking from an individualist perspective consistent with the first two meta-axioms (**D** and **S**). John Maynard Keynes was another such thinker. The difference was that he was not a believer! Indeed, his best work reflects the '*We are damned if we know*' logic. In short, Keynes did not believe in the *inevitability* of *any* kind of *regularity*; of either the equilibrium (**E**) or the spontaneous order (**O**) types. For this reason,

Table 13.1

awards him only two ticks, courtesy of his roots in his teachers' (and in particular Alfred Marshall's) marginalism.

4

I end this discussion of

Table 13.1

with the last column which captures whether the 'mind-frame' of the thinkers in each different row is predicated upon human reductionism; upon, that is, a readiness to think of men and women, indeed of children too, as analytically equivalent to machines, to a mathematical mapping of outcomes to some index of preference satisfaction, or, at best, to algorithms of sorts.

The British classical economists embraced human reductionism clearly and knowingly. Adam Smith and David Ricardo left no room in their political

economics for economic insights that are uniquely due to the indeterminacy of human nature. In their economic writings, humans appear as machine-like, preprogrammed creatures.

5

The first political economist to have based an important *economic insight* on the irreducibility of the human person to a quantifiable, machine-like entity, was of course Karl Marx. But so did the Austrians and, of course, Keynes (thus the crosses in the respective cells in the last column).

Table 13.1

The six meta-axioms of political economics

	Agency		Regularity		Reducibility of human action (R)
	Strong Meth. Ind. (D)	Strong Meth. Instr. (S)	Strong Meth. Equilibration (E)	Spontaneous Order (O)	
Smith–Ricardo–Sraffa	×	×	✓	×	✓
Marx	×	×	✓	×	×
Marginalists	✓	✓	×	×	✓
Marginalists-cum- Neoclassicists	✓	✓	✓	×	✓
Austrian Marginalists	✓	✓	×	✓	×
Keynes	✓	✓	×	×	×
	×	×	×	×	×

The Austrians rejected the idea that information equals knowledge and that it is a technical matter to aggregate it all in one large hard disk-like device. They rejected the notion of some economy-wide equilibrium because they rejected the idea that human knowledge is like grains of sand to be piled up by a process of mechanical aggregation. Similarly, Keynes opposed the view that investors and consumers predict the future in a manner ontologically no different to performing a technically difficult computation. For reasons that are related to their appreciation of the irrepressible nature of indeterminacy, both the Austrians and Keynes thought that there is no such thing as a sufficiently narrow set of rational expectations that agents, if clever enough, could home in on.

In summary, Karl Marx, the Austrians, and John Maynard Keynes set themselves apart from the rest of the political economists by treating the indeterminate human element as a crucial analytical datum. Of these three, however, only Keynes felt sufficiently liberated from his own ideological imperative to present an argument in favour of, or against, capitalism. He took it for granted that he liked capitalism and did not need to prove its superiority or desirability. What concerned him was capitalism's capacity for self-suicide. Period.

In this spirit, Keynes embarked upon his *General Theory* in order to furnish practical advice on how to manage capitalism's depressive character effectively. For this reason

Table 13.1

's penultimate row (dedicated to Keynes) features no ticks in the last three columns: Keynes, having rejected that human reasoning can be reduced to the operations of an algorithm, did not trust capitalism to equilibrate or regulate itself.

Which brings us to

Table 13.1

's last row, the one enigmatically left un-labelled. This is 'my' column and its purpose is to act as a brief manifesto. It is a simple four word manifesto: *No meta-axioms please*. 'Closed' models are destined to fall prey to indeterminacy's voracious appetite. The only scientific truth about capitalism is precisely its *radical indeterminacy*, a condition which makes it impossible to use science's tools (e.g. calculus and statistics) to second guess it. The more we feel we have capitalism's number, the closer we get to the moment when it will astonish us with (what our 'closed' models told us was) an almost zero probability event. When the improbable becomes fact, our only hope is that the casualties will not be too numerous.

But what are the sources of the *radical indeterminacy*? Keynes answered that question partially. In multi-sector, financialised capitalist economies, consumers and investors lack the data that would allow them, even if they possessed God's own

computing capacities, to construct a determinate mathematical expectation of what the future holds. Like ships with de-magnetised compasses sailing in a starless night, they tend to follow one another along self-confirming paths. *Even*

if captained by supremely experienced sailors, they may make it safely to port or they may all be led astray, ending up marooned on shoals from which they cannot extricate themselves.

In summary, because of the impossibility of uniquely rational answers to pressing questions such as 'How much should I save?' and 'Should I invest now?', consumption and investment are at the mercy of the *Cunning of Reason* (which Keynes mislabelled *animal spirits*). But there is another source of *radical indeterminacy* that Keynes ignored, possibly because he was unwilling to recognise its location in the veins of a class of people whom he was conditioned to look down upon: Human *labour* which (as Karl Marx taught us) is the life-giving force that runs through capitalism bestowing value and even life upon mere 'things', albeit only as long as it remains *indeterminate; irreducible*, that is, *to an electricitylike force*. It is this vivifying, indeterminate energy that creates *capital* out of mere machines; a relatively newfangled force with the astonishing capacity both to liberate and to enslave the humans that work it and the humans that own it alike.

In brief, without a grasp of the *dialectical nature of both labour and capital* it becomes impossible to understand:

- (a) the dynamics of a capitalist economy, *and*
- (b) the ways in which irrepressibly free humans become increasingly enslaved by their artefacts.

My hypothesis here is that to make sense of capitalism we need to capture (a) and (b), and to combine them with Keynes' successful escape from determinate models of investment and aggregate demand. The task is equivalent to introducing into political economics, as 'data', the two sources of *radical indeterminacy*:

- (i) the irreducibility of labour input, and thus capital, to some well defined metric;
and
- (ii) the irreducibility of human forecasts to a well defined mathematical expectation function.

As long as (i) and (ii) are combined with a determination to assume neither *equilibrium* (which was Marx's error) nor *spontaneous order* (as is the Austrians' religious wont), we stand a chance of grasping our present moment in history. Moreover, the events of 2008 are better understood as our collective punishment for the economists' greatest sin: the assumption that *radical indeterminacy* can be tamed by means of formalist meta-axioms at one level and simple pricing formulae at another level, like the ones which financial engineering used to procure its splendid fantasies.

13.6 Postscript

This book was about failure and power, as I promised in the Preface. Its main theme was about the economics' profession grandest achievement: Of the continual conversion of theoretical flops into untold social, political and economic power. Power for the economists themselves but also power for the politicians (who use the economists' models to pass their toxic policies through bamboozled Parliaments and cabinet meetings) and for the financiers (who also use the same models in order to extend an air of legitimacy to their toxic 'products').

Meanwhile, the book's subtitle made it clear that it constitutes a kind of analytical auto-biography, with a loftier aim firmly attached to any delusion of grandeur: to warn, that is, newcomers to the economics profession, particularly young graduates, that their chosen profession is riddled with booby-traps and landmines. That they will be judged

not on the basis of establishing the truth-status of their theories but on how efficiently and intelligently they reproduce models based on meta-axioms whose business it is to obfuscate capitalism's true nature and workings.

Lastly, in this concluding chapter, the reader will have noticed that I chose theatre as my proxy to social life and as a means of recasting, and summing up, my critique of economics. The dramatic hollowness that must plague a theatrical production modelled along neoclassical lines is the mirror image of the irrelevance of the economists' models of really-existing capitalism. As practitioners of economics who seek to escape from the latter's institutionalised misanthropy, we can do worse than continue to aspire to an economic narrative that contains some of the panache of a play by Sophocles or Shakespeare. Otherwise, we shall remain irreversibly immersed in a variety of economics more reminiscent of John Wayne movies, contributing significantly until well after we retire to the suffering of so many people with fewer opportunities to expose the ways of financialised capitalism than we were granted.

Notes

- 1 My idea of imagining how a neoclassicist playwright would go about her business of writing her play first occurred to me when I was writing my 1991 book *Rational Conflict*. See pp. 196–200.
- 2 See Varoufakis *et al.* (2011).
- 3 See Varoufakis (1991), Varoufakis (1998) and Hargreaves-Heap and Varoufakis (2004).
- 4 These are, naturally, broad brushstrokes by which to paint the portrait of major intellectuals. One might plausibly argue, for instance, that by the time Keynes had finished his *General Theory* very few of his roots in Marshall's marginalism remained, at least when thinking of the macroeconomy. In this sense, the two ticks in Keynes' row ought to be fainter than the corresponding ticks in the rows of the Austrians or the Marginalists.
- 5 This is not to say that they did not acknowledge the special features of human nature in other writings. Adam Smith, for instance, did so extensively in his *Moral Sentiments*. Our point here is that human labour and decision making is rendered mechanistic in their writings on political economics.

References

- Ades, A. and R. Di Tella (1996). 'Rents, competition, and corruption', *American Economic Review*, 89, 982–93.
- Akerlof, G. (1980). 'A theory of social custom of which unemployment may be one consequence', *Quarterly Journal of Economics*, XCIV, 749–75.
- Akerlof, G. (1982). 'Labor contracts as partial gift exchange', *Quarterly Journal of Economics*, 97, 543–69.
- Akerlof, G. (2007). 'The missing motivation of macroeconomics', *American Economic Review*, 97, 5–36.
- Allais, M. (1953). 'Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'Ecole Americaine', *Econometrica*, 21, 503–46.
- Andreoni, J. (1990). 'Impure altruism and donations to public goods: A theory of warm-glow giving', *Economic Journal*, 100, 464–77.
- Aristotle (1987). *Nicomachean Ethics*, transl. by J. Welson. New York: Prometheus.
- Arnsperger, C. and Y. Varoufakis (1999). 'Solidarity and rational contemplation', Mimeo, University of Sydney, Department of Economics, May.
- Arnsperger, C. and Y. Varoufakis (2003). 'Toward a theory of solidarity', *Erkenntnis*, 59, 157–88.
- Arrow, K. (1959). 'Towards a theory of price adjustment', in M. Abramovitz (ed.), *Allocation of Economic Resources*. Stanford, CA: Stanford University Press.
- Arrow, K. (1994). 'Methodological individualism and social knowledge', *American Economic Review (Papers and Proceedings)*, 84, 1–9.
- Arrow, K. (2009). 'Some developments in economic theory since 1940: An eyewitness account', *Annual Review of Economics*, 1, 1–16 (see www.econ.annualreviews.org).
- Arrow, K. and G. Debreu (1954). 'Existence of an Equilibrium for a competitive economy', *Econometrica*, 22, 265–90.
- Asdigan, N., E. Cohn, and M. Blum, (1994). 'Gender differences in distributive justice: The role of self-representation revisited'. *Sex Roles*, 30, 303–18.
- Ashenfelter, O. and G. E. Johnson (1969). 'Bargaining theory, trades unions and industrial strike activity', *American Economic Review*, 59, 35–49.
- Aspromourgos, T. (1986). 'On the origins of the term neoclassical', *Cambridge Journal of Economics*, 10, 265–70.
- Aumann, R. (1976). 'Agreeing to disagree', *Annals of Statistics*, 4, 1236–9.
- Aumann, R. (1987). 'Correlated equilibrium as an expression of Bayesian rationality', *Econometrica*, 55, 1–18.
- Auriol, E. (2006). 'Corruption in procurement and public purchase', *International Journal of Industrial Organization*, 24, 867–85.
- Axelrod, R. and D. Dion (1988). 'The further evolution of cooperation', *Science*, 242, 1385–91.
- Babcock, L. and G. Loewenstein (1997). 'Explaining bargaining impasse: The role of self-serving biases', *Journal of Economic Perspectives*, 11, 109–26.
- Babcock, L., G. Loewenstein, S. Issachoroff, and C. Camerer (1995). 'Biased judgments of fairness in bargaining', *American Economic Review*, 85, 1337–43.
- Bacharach, M. (1987a). 'A theory of rational decision in games', *Erkenntnis*, 27, 17–55.
- Bacharach, M. (1987b). '"We" equilibria: A variable frame theory of co-operation', Mimeo.
- Bacharach, M. (1999). 'Interactive team reasoning: A contribution to the theory of cooperation', *Research in Economics*, 53, 117–47.
- Bacharach, M. and M. Bernasconi (1997). 'The variable frame theory of focal points: An experimental study', *Games and Economic Behaviour*, 19, 1–45.
- Becker, G. (1974). 'A theory of social interactions', *Journal of Political Economy*, 82, 1063–93.
- Becker, G. (1976). *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.
- Benhabib, S. (1984). 'Obligation, contract and exchange: On the significance of Hegel's abstract right', in Z. Pelczynski (ed.), *The State and Civil Society: Studies in Hegel's Political Philosophy*. Cambridge: Cambridge University Press.
- Bergin, J. and B. L. Lipman (1996). 'Evolution with state-dependent mutations', *Econometrica*, 64, 943–56.
- Berlin, I. (1958). 'Two concepts of liberty', reprinted in *Four Essays on Liberty*. Oxford: Oxford University Press.
- Bernheim, D. (1984). 'Rationalisable strategic behaviour', *Econometrica*, 52, 1007–28.
- Bini, P. and L. Bruni (1998). 'Intervista a Gerard Debreu', *Storia del Pensiero Economico*, 35, 3–29.
- Binmore, K. (1987). 'Modeling rational players: Part I', *Economics and Philosophy*, 3, 179–214.
- Binmore, K. (1988). 'Modeling rational players: Part II', *Economics and Philosophy*, 4, 9–55.
- Binmore, K. (1998). *Just Playing*. Cambridge, MA: MIT Press.
- Binmore, K. and L. Samuelson (1993). 'Learning to play the ultimatum game', Mimeo.
- Binmore, K., A. Rubinstein, and A. Wolinsky (1986). 'The Nash bargaining solution in economic modelling', *Rand Journal of Economics*, 17, 176–88.
- Binmore, K., J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked (2002). 'A backward induction experiment', *Journal of Economic Theory*, 104, 48–88.
- Bishop, R. L. (1984). 'A Hicks-Zeuthen theory of bargaining', *Econometrica*, 52, 410–17.
- Blair, D. and D. Crawford (1984). 'Labor union objectives and collective bargaining', *Quarterly Journal of Economics*, 99, 547–66.
- Blanchflower, D., N. Millward, and A. Oswald (1991). 'Unionism and employment behaviour', *The Economic Journal*, 101, 815–35.
- Blaug, M. (1992). *The Methodology of Economics: Or How Economists Explain*, 2nd edn. Cambridge: Cambridge University Press.
- Blaug, M. (1997). 'Ugly currents in modern economics', Plenary Sessions Presentation, *Fact or Fiction? Perspectives on Realism and Economics*, Rotterdam, 14–15 November.
- Bliss, C. (2005). 'Introduction. The theory of capital: A personal overview', in C. Bliss, A. Cohen, and G. C. Harcourt (eds),

- Capital Theory*, Vol. 1. Cheltenham: Edward Elgar.
- Bliss, C. and R. Di Tella (1997). 'Does competition kill corruption?', *Journal of Political Economy*, 105, 1001–23.
- Bowles, S. (1998). 'Endogenous preferences: The cultural consequences of markets and other economic institutions', *Journal of Economic Literature*, 36, 75–111.
- Brennan, G. and P. Pettit (2000). 'The hidden economy of esteem', *Economics and Philosophy*, 16, 77–98.
- Brown, J. and O. Ashenfelter (1986). 'Testing the efficiency of employment contracts', *Journal of Political Economy*, 94, S40–S80.
- Burmeister, E. (2000). 'The capital theory controversy', in H. Kurz (ed.), *Critical Essays on Piero Sraffa's Legacy in Economics*. Cambridge: Cambridge University Press.
- Camerer, C. (1997). 'Progress in behavioral game theory', *Journal of Economic Perspectives*, 11, 167–88.
- Camerer, C. and H. Thaler (1995). 'Anomalies: Ultimatum, dictators and manners', *Journal of Economic Perspectives*, 9, 209–19.
- Camerer, C. and T.-H. Ho (1999). 'Experience-weighted attraction learning in normal form games', *Econometrica*, 67, 827–74.
- Card, D. (1986). 'Efficient contracts with costly adjustment: Short-run employment determination for airline mechanics', *American Economic Review*, 76, 1045–71.
- Carling, A. (1990). 'In defence of Rational Choice: A reply to Ellen Meiskins Wood', *New Left Review*, 184, 97–109.
- Carmichael, H. and W. MacLeod (1997). 'Fair territory: Preferences, bargaining and the endowment effect', mimeo.
- Chapman, B. (1998). 'More easily done than said: Rules, reasons, and rational social choice', *Oxford Journal of Legal Studies*, 18, 293.
- Coase, R. (1994). *Essays on Economics and Economists*. Chicago: Chicago University Press.
- Cohen, A. and G. Harcourt (2003). 'Whatever happened to the Cambridge capital theory controversies?', *Journal of Economic Perspectives*, 17, 199–214.
- Colander, D. (2005a). 'The making of an economist redux', *Journal of Economic Perspectives*, 19, 175–98.
- Colander, D. (2005b). 'The future of economics: The appropriately educated in pursuit of the knowable', *Cambridge Journal of Economics*, 29, 927–41.
- Colander, D., R. Holt, and J. Rosser Jr. (2004a). *The Changing Face of Economics: Interviews with Cutting Edge Economists*. Ann Arbor: University of Michigan Press.
- Colander, D., R. Holt, and J. Rosser Jr. (2004b). 'The changing face of economics', *Review of Political Economy*, 16, 485–99.
- Coleman, J. (1990). *Foundations of Social Theory*. Cambridge: Harvard University Press.
- Cross, J. (1969). *The Economics of Bargaining*. New York: Basic Books.
- Dahlberg, M. and E. Mork (2006). 'Public employment and the double role of bureaucrats', *Public Choice*, 126, 387–404.
- Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
- Dasgupta, P. (2002). 'Modern economics and its critics', in U. Mäki (ed.), *Fact and Fiction in Economics: Models, Realism and Social Construction*. Cambridge: Cambridge University Press.
- Davis, J. (2003). *The Theory of the Individual in Economics: Identity and Value*. London and New York: Routledge.
- Davis, J. (2006). 'The turn in economics: Neoclassical dominance to mainstream pluralism?', *Journal of Institutional Economics*, 2, 1–20.
- Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.
- Dawkins, R. (1980). 'Good strategy or evolutionarily stable strategy?', in G. W. Barlow and J. Silverberg (eds), *Sociobiology: Beyond Nature/Nurture?* Boulder, CO: Westview Press.
- Debreu, G. (1959). *Theory of Value: An Axiomatic Study of Economic Equilibrium*. New York: Wiley.
- Debreu, G. (1974). 'Excess demand functions', *Journal of Mathematical Economics*, 1, 15–21.
- Debreu, G. (1986). 'Theoretic models: Mathematical form and economic content', *Econometrica*, 54, 1259–70.
- Dekel, E. and S. Scotchmer (1992). 'On the evolution of optimizing behaviour', *Journal of Economic Theory*, 57, 392–406.
- Derrida, J. (1973). *Speech and Phenomena and Other Essays on Husserl's Theory of Signs*. Evanston, IL: Northwestern University Press.
- Derrida, J. (1978). *Writing and Difference*. London: Routledge & Kegan Paul.
- Diamond, J. (1996). *Guns, Germs and Steel: The Fate of Human Societies*. New York: Norton.
- Doiron, D. (1992). 'Bargaining power and wage-employment contracts in a unionized industry', *International Economic Review*, 33, 583–606.
- Dow, S. C. (1995). 'The appeal of neoclassical economics: Some insights from Keynes's epistemology', *Cambridge Journal of Economics*, 19, 715–33.
- Dunlop, J. (1944). *Wage Determination under Trade Unions*. New York: Macmillan.
- Dutta, P. K. (1999). *Strategies and Games: Theory and Practice*. Cambridge, MA: MIT Press.
- Ellsberg, D. (1956). 'Theory of the reluctant duellist', *American Economic Review*, 46, 909–23.
- Ellsberg, D. (1961). 'Risk, ambiguity and the Savage axioms', *Economic Journal*, 64, 643–69.
- Elster, J. (1982). 'Marxism, functionalism and game theory', *Theory and Society*, 11, 453–82.
- Elster, J. (ed.) (1986a). *Rational Choice*. Cambridge: Cambridge University Press.
- Elster, J. (1986b). *Making Sense of Marx*. Cambridge: Cambridge University Press.
- Elster, J. (1989). 'Social norms and economic theory', *Journal of Economic Perspectives*, 3, 99–117.
- Emerson, M. (2002). 'Corruption and industrial dualism in less developed countries', *Journal of International Trade and Economic Development*, 11, 63–76.
- Emerson, P. (2006). 'Corruption, competition and democracy', *Journal of Development Economics*, 81, 193–212.
- England, P. (1993). 'The separative self: Androcentric bias in neoclassical assumptions', in M. A. Ferber and J. A. Nelson (eds), *Beyond Economic Man: Feminist Theory and Economics*. Chicago: University of Chicago Press.
- Erev, I. and A. Roth (1998). 'Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria', *American Economic Review*, 88, 848–81.

- Erev, I., Y. Bereby-Meyer, and A. Roth (1999). 'The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models', *Journal of Economic Behavior & Organization*, 39, 111–28.
- Farber, H. S. (1976). 'Bargaining theory, wage outcomes, and the occurrence of strikes: An econometric analysis', *American Economic Review*, 68, 262–71.
- Fehr, E. and S. Gächter (2000). 'Fairness and retaliation: The economics of reciprocity', *Journal of Economic Perspectives*, 14, 159–81.
- Fine, B. (2008). 'The general impossibility of neoclassical economics: Or does Bertrand Russell deserve a Nobel Prize for Economics?', mimeo.
- Finelli, R. (1990). 'Production of commodities and production of images: reflection on modernism and postmodernism', mimeo, Faculty of Philosophy, University of Rome.
- Foddy, M. (1989). 'Information control as a bargaining tactic in social exchange', in E. J. Lawler and B. Markovsky Barry (eds), *Advances in Group Processes*, Vol. 6. Greenwich, CT: JAI Press.
- Foster, D. and H. P. Young (1990). 'Stochastic evolutionary game dynamics', *Theoretical Population Biology*, 38, 219–32.
- Foucault, M. (1967). *Madness and Civilisation*, London: Tavistock.
- Frankfurt, H. (1971). 'Freedom of the will and the concept of reason', *Journal of Philosophy*, 68, 5–20.
- Friedman, D. (1991). 'Evolutionary games', *Econometrica*, 59, 637–66.
- Friedman, D. (1996). 'Equilibrium in evolutionary games: Some experimental results', *Economic Journal*, 106, 1–25.
- Friedman, M. (1962). *Free to Choose*. Melbourne: Macmillan.
- Fudenberg, D. and C. Harris (1992). 'Evolutionary dynamics in games with aggregate shocks', *Journal of Economic Theory*, 57, 420–41.
- Fudenberg, D. and J. Tirole (1983). 'Sequential bargaining with incomplete information', *Review of Economic Studies*, 50, 221–41.
- Fudenberg, D., D. Kreps, and A. Levine (1988). 'On the robustness of equilibrium refinements', *Journal of Economic Theory*, 44, 533–44.
- Fullbrook, E. (2003). *The Crisis in Economics. The Post-Autistic Economics Movement: The First 600 Days*. London: Routledge.
- Fullbrook, E. (2004). *A Guide to What's Wrong with Economics*. London: Anthem Press.
- Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.
- Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). 'Psychological games and sequential rationality', *Games and Economic Behavior*, 1, 60–79.
- Gilbert, M. (1989). *On Social Facts*. London: Routledge.
- Goodwin, R. (1967). 'A growth cycle', in C. Feinstein (ed.), *Socialism, Capitalism and Economic Growth*. Cambridge: Cambridge University Press.
- Gul, F. and H. Sonnenschein (1988). 'On delay in bargaining with one-sided uncertainty', *Econometrica*, 601–11.
- Habermas, J. (1990). *Moral Consciousness and Communicative Action*, transl. by C. Lenhardt and S. Nicholsen. Oxford: Polity Press.
- Halevi, J. (1998). 'Capital and growth: Its relevance as a critique of neo-classical and classical economic theories', *Indian Journal of Applied Economics*, 7, 79–98.
- Harcourt, G. (1972). *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.
- Hargreaves-Heap, S. (1989). *Rationality in Economics*. Oxford: Basil Blackwell.
- Hargreaves-Heap, S. and Y. Varoufakis (1995). *Game Theory: A Critical Introduction*. London and New York: Routledge.
- Hargreaves-Heap, S. and Y. Varoufakis (2002). 'Some experimental results on the evolution of discrimination, co-operation and perceptions of fairness', *Economic Journal*, 112, 678–702.
- Hargreaves-Heap, S. and Y. Varoufakis (2004). *Game Theory: A Critical Text*. London and New York: Routledge.
- Harris, L. (1978). 'Catastrophe theory, utility theory and the animal spirits expectations', *Australian Economic Papers*, 19, 268–82.
- Harrison, E. (2006). 'Unpacking the anti-corruption agenda: Dilemmas for anthropologists', *Oxford Development Studies*, 34, 15–29.
- Harsanyi, J. (1973). 'Games with randomly disturbed payoffs: A new rationale for mixed strategies', *International Journal of Game Theory*, 2, 1–23.
- Harsanyi, J. and R. Selten (1972). 'A generalised Nash solution for two-person bargaining games with incomplete information', *Management Science*, 18, 80–106.
- Hart, O. (1989). 'Bargaining and strikes', *Quarterly Journal of Economics*, 104, 25–44.
- Hayes, B. (1984). 'Unions and strikes with asymmetric information', *Journal of Labor Economics*, 2, 57–83.
- Hegel, G. W. F. (1931). *The Phenomenology of Mind*, transl. by J. Baillie. London: Macmillan.
- Hegel, G. W. F. (1942). *Philosophy of Right*, transl. by T. Knox. Oxford: Clarendon Press.
- Hegel, G. W. F. (1953). *Reason in History*, transl. by R. Hartman. New York: Library of Liberal Arts.
- Heiding, H. and H. Moulin (1991). 'The solidarity axiom in parametric surplus-sharing problems', *Journal of Mathematical Economics*, 20, 249–70.
- Hewitson, G. (1999). *Feminist Economics: Interrogating the Masculinity of Rational Economic Man*. Cheltenham: Edward Elgar.
- Hicks, J. R. (1966, first edition 1932). *The Theory of Wages*. London: Macmillan.
- Hobbes, T. (1991). *Leviathan*, ed. by R. Tuck. Cambridge: Cambridge University Press.
- Hodgson, G. M. (1993). *Economics and Evolution: Bringing Life Back into Economics*. Oxford: Polity.
- Hodgson, G. (1999). *Evolution and Institutions: On Evolutionary Economics and the Evolution of Economics*. Cheltenham: Edward Elgar.
- Hodgson, G. (2007). 'Meanings of methodological individualism', *Journal of Economic Methodology* 14, 211–26.
- Hoffman, E., K. McCabe, and V. Smith (1996). 'Social distance and Other-regarding behavior in dictator games', *American Economic Review*, 86, 653–60.
- Hollis, M. (1987). *The Cunning of Reason*. Cambridge: Cambridge University Press.

- Hollis, M. (1989). 'Honour among thieves', *Proceedings of the British Academy*, LXXV, 163–80.
- Hollis, M. (1990). 'Moves and motives in the games people play', *Analysis*, 50, 49–62.
- Hollis, M. (1991). 'Penny-pinching and backward induction', *Journal of Philosophy*, 89, 472–88.
- Hollis, M. (1993). 'The agriculture of the mind', in D. Gauthier and R. Sugden (eds), *Rationality, Justice and the Social Contract*. Hemel Hempstead: Wheatsheaf.
- Hollis, M. (1998). *Trust within Reason*. Cambridge: Cambridge University Press.
- Hollis, M. and R. Sugden (1993). 'Rationality in action', *Mind*, 102, 1–35.
- Hume, D. (1888). *Treatise of Human Nature*, ed. by L. A. Selby-Bigge. Oxford: Oxford University Press.
- Hurley, S. (1989). *Natural Reasons*. Oxford: Oxford University Press.
- Kahn, A., V. O'Leary, J. Krulewitz, and H. Lamm, (1980). 'Equity and equality: Male and female means to a just end', *Basic and Applied Social Psychology*, 1, 173–97.
- Kandori, M., G. Mailath, and R. Rob (1993). 'Learning, mutation, and long run equilibria in games', *Econometrica*, 61, 29–56.
- Kant, I. (1781, 2003). *Critique of Pure Reason*, transl. by N. K. Smith. London: Palgrave Macmillan.
- Kant, I. (1788, 1949). *Critique of Practical Reason* in *Critique of Practical Reason and Other Writings*, transl. and ed. by L. W. Beck. Cambridge: Cambridge University Press.
- Kant, I. (1959). *The Fundamental Principles of the Metaphysic of Morals*. London: Longmans.
- Kennan, J. (1987). 'The economics of strikes', in O. Ashenfelter and R. Layard (eds), *Handbook of Labour Economics*. Amsterdam: North Holland.
- Kennan, J. and R. Wilson (1989). 'Strategic bargaining models and the interpretation of bargaining models', *Journal of Applied Econometrics* 4, S87–S130.
- Kirman, A. (1989). 'The intrinsic limits of modern economic theory: The emperor has no clothes', *Economic Journal*, 99, 126–39.
- Kirsch, J. (1966). *Shakespeare's Royal Self*. New York: Putnam.
- Kreps, D. (1990). *Game Theory and Economic Modeling*. New York: Oxford University Press.
- Kreps, D. and R. Wilson (1982). 'Reputation and imperfect information', *Journal of Economic Theory*, 27, 253–79.
- Kreps, D., P. Milgrom, J. Roberts, and R. Wilson (1982). 'Rational cooperation in the finitely repeated prisoner's dilemma', *Journal of Economic Theory*, 27, 245–52.
- Kurz, H. and N. Salvadori (1993). 'Von Neumann's growth model and the "classical" tradition', *European Journal of the History of Economic Thought*, 1, 130–60.
- Lawson, T. (1997). *Economics and Reality*. London and New York: Routledge.
- Lawson, T. (2003). *Reorienting Economics*. London and New York: Routledge.
- Leijonhufvud, A. (1968). *Keynesian Economics and the Economics of Keynes*. Cambridge: Cambridge University Press.
- Leontief, W. (1946). 'The pure theory of the guaranteed annual wage contract', *Journal of Political Economy*, 54, 76–9.
- Levi-Strauss, C. (1966). *The Savage Mind*. London: Weidenfeld and Nicolson.
- Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
- Lyotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge*. Manchester: Manchester University Press.
- McCloskey, D. (1985). *The Rhetoric of Economics*. Madison: University of Wisconsin Press.
- McConnell, S. (1989). 'Strikes, wages and private information', *American Economic Review*, 89, 801–15.
- McCurdy, T. and J. Pencavel (1986). 'Testing between competing models of wage and employment determination in unionised markets', *Journal of Political Economy*, 94, S3–S39.
- McDaniell, T., E. Rutström, and M. Williams (1994). 'Incorporating fairness into game theory and economics: An experimental test with incentive compatible belief elicitation', mimeo.
- Macdonald, I. and R. Solow (1981). 'Wage bargaining and employment', *American Economic Review*, 71, 896–908.
- Machin, S. and S. Wadhvani (1991). 'The effects of unions on organisational change and employment', *Economic Journal*, 101, 835–54.
- McKelvey, R. and T. Palfrey (1992). 'An experimental study of the centipede game', *Econometrica*, 60, 803–36.
- McPherson, C. B. (1973). *Democratic Theory: Essays in Retrieval*. Oxford: Clarendon Press.
- Mailath, G. (1998). 'Do people play Nash equilibrium? Lessons from evolutionary game theory', *Journal of Economic Literature*, 36, 1347–74.
- Mailath, G. and A. Postlewaite (1990). 'Workers versus firms: Bargaining over a firm's value', *Review of Economic Studies*, 57, 369–80.
- Major, B. and J. Adams (1983). 'Role of gender, inter-personal orientation and self-representation in distributive justice behavior', *Journal of Personality and Social Psychology*, 45, 598–608.
- Major, B., W. Bylsma, and C. Cozzarelli (1989). 'Gender differences in distributive justice preferences: The impact of domain', *Sex Roles*, 21, 487–97.
- Manning, A. (1987). 'An integration of trade union models in a sequential bargaining framework', *Economic Journal*, 97, 121–39.
- Mantel, R. (1974). 'On the characterization of aggregate excess demand', *Journal of Economic Theory*, 7, 348–53.
- Margolis, H. (1981). 'A new model of Rational Choice', *Ethics*, 91, 265–79.
- Marshall, A. (1891). *Principles of Economics*. London: Macmillan.
- Marx, K. (1963). *The Poverty of Philosophy*. New York: International Publishers.
- Marx, K. (1964). *Early Writings*. New York: McGraw-Hill.
- Marx, K. (1972). *Capital I–III*. London: Lawrence and Wishart.
- Marx, K. (1974). *Grundrisse*. New York: McGraw-Hill.
- Marx, K. and F. Engels (1979). *Collected Works*. London: Lawrence and Wishart.
- Matsui, A. (1992). 'Best response dynamics and socially stable strategies', *Journal of Economic Theory*, 57, 343–62.
- Mauro, P. (1995). 'Corruption and growth', *Quarterly Journal of Economics*, 110, 681–712.
- Maynard Smith, J. and G. Price (1974) 'The theory of games and the evolution of animal conflict', *Journal of Theoretical Biology*, 47, 209–21.

- Mehta, J., C. Starmer, and R. Sugden (1994). 'The nature of salience: An experimental investigation of pure coordination games', *American Economic Review*, 84, 658–73.
- Midgley, M. (1994). *The Ethical Primate: Humans, Freedom and Morality*. London: Routledge.
- Mirowski, P. (1989). *More Heat Than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge: Cambridge University Press.
- Mirowski, P. and E. R. Weintraub (1994). 'The pure and the applied: Bourbakism comes to mathematical economics', *Science in Context*, 7, 245–72.
- Nash, J. (1950). 'The bargaining problem', *Econometrica*, 18, 155–62.
- Nash, J. (1951). 'Non-cooperative games', *Annals of Mathematics*, 54, 286–95.
- Nash, J. (1953). 'Two person cooperative games', *Econometrica*, 21, 128–40.
- Naylor, R. (1987). 'Strikes, free-riders and social customs', *Quarterly Journal of Economics*, 104, 771–83.
- Nickell, S. and M. Andrews (1983). 'Trade unions, real wages and employment in Britain: 1951–1979', *Oxford Economic Papers*, 35, 183–206.
- Nietzsche, F. (1956). *Genealogy of Morals*. New York: Doubleday.
- Nietzsche, F. (1964). 'On truth and falsity in their ultramoral sense', in O. Levy, *The Complete Works of Friedrich Nietzsche*. New York: Russell and Russell.
- Nietzsche, F. (1973). *Beyond Good and Evil: Prelude to a Philosophy of the Future*. Hammondsworth: Penguin.
- Norris, C. (1985). *The Contest of Faculties: Philosophy and Theory after Deconstruction*. London: Methuen.
- Nowak, A. S. and T. Radzik (1994). 'A solidarity value for n -person transferable utility games', *International Journal of Game Theory*, 23, 43–8.
- Nozick, R. (1974). *Anarchy, State and Utopia*. New York: Basic Books.
- Olson, M. (1965). *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.
- Oswald, A. (1983). 'Wage bargains are on the labour demand curve', mimeo, Princeton University.
- Oswald, A. (1994). 'Efficient contracts are on the labour demand curve: Theory and facts', *Labour Economics*, 1, 85–113.
- Patokos, T. (2005). 'On the evolutionary fitness of bounded rationality heterogeneous populations in antagonistic interactions', *American Journal of Applied Sciences*, 2, 61–72.
- Peirce, C. S. (1932). *Collected Papers*, Vol. 2. Cambridge, MA: Harvard University Press.
- Pettit, F. and R. Sugden (1989). 'The paradox of backward induction', *Journal of Philosophy*, 86, 169–82.
- Polanyi, K. (1945, 1957). *Primitive, Archaic and Modern Economies*. London: Routledge & Kegan Paul.
- Poston, T. and I. Stewart (1978). *Catastrophe Theory and Its Applications*. London: Pitman.
- Rabin, M. (1993). 'Incorporating fairness into economics and game theory', *American Economic Review*, 83, 1281–302.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Reny, P. (1993). 'Backward induction, normal form perfection and explicable equilibria', *Econometrica*, 60, 627–49.
- Roemer, J. (1985). 'Rationalizing revolutionary ideology', *Econometrica*, 53, 85–108.
- Roemer, J. (ed.) (1986). *Analytical Marxism*. Cambridge: Cambridge University Press.
- Rouille d'Orfeuil, H. (2002). *Finances et Solidarité*. Paris: La Découverte.
- Rousseau, J.-J. (1973). *The Social Contract, Discourses*, ed. by G. Cole. London: Dent.
- Rowe, C. (1993). 'Ethics in Ancient Greece', in P. Singer (ed.), *A Companion to Ethics*, Oxford: Blackwell.
- Rubinstein, A. (1982). 'Perfect equilibrium in a bargaining model', *Econometrica*, 50, 97–109.
- Rubinstein, A. (1985). 'A bargaining model with incomplete information about preferences', *Econometrica*, 53, 1151–72.
- Ryan, M. (1982). *Marxism and Deconstruction: A Critical Articulation*. Baltimore, MD: Johns Hopkins University Press.
- Saha, B. and T. Thampy (2006). 'Extractive bribe and default in subsidized credit programs', *Journal of Economic Behavior and Organization*, 60, 182–204.
- Savage, L. (1954). *The Foundations of Statistics*. New York: Wiley.
- Schelling, T. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
- Schotter, A., A. Weiss, and I. Zapater (1996). 'Fairness and survival in ultimatum and dictatorship games', *Journal of Economic Behavior & Organization*, 31, 37–56.
- Schumpeter, J. (1908). *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Munich and Leipzig: Duncker and Humblot.
- Selten, R. and A. Ockenfels (1998). 'An experimental solidarity game', *Journal of Economic Behavior and Organization*, 34, 517–39.
- Sen, A. (1970). 'The impossibility of a Paretian Liberal', *Journal of Political Economy*, 78, 152–7.
- Sen, A. (1974). 'Choice, orderings and morality', in S. Korner (ed.), *Practical Reason*. Oxford: Basil Blackwell.
- Sen, A. (1977). 'Rational fools: A critique of the behavioral foundations of economic theory', *Philosophy and Public Affairs*, 6, 317–44.
- Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.
- Shleifer, A. and R. Vishny (1993). 'Corruption', *Quarterly Journal of Economics*, 104, 537–64.
- Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.
- Smith, A. (1976). *The Theory of Moral Sentiments*, ed. by D. Raphael and A. Macfie. Oxford: Clarendon Press.
- Sonnenschein, H. (1972). 'Market excess demand functions', *Econometrica*, 40, 549–63.
- Sonnenschein, H. (1973). 'Do Walras' identity and continuity characterize the class of community excess demand functions?', *Journal of Economic Theory*, 6, 345–54.
- Sprumont, Y. (1996). 'Axiomatizing ordinal welfare egalitarianism when preferences may vary', *Journal of Economic Theory*, 68, 77–100.
- Sraffa, P. (1975). *The Production of Commodities by Means of Commodities: Prelude to a Critique of Political Economy*. Cambridge: Cambridge University Press.
- Starmer, C. (2000). 'Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk', *Journal of Economic Literature*, 38, 332–82.
- Stiglitz, J. (2002). 'There is no invisible hand', *The Guardian*, 20 December.
- Sugden, R. (1982). 'On the economics of philanthropy', *Economic Journal*, 92, 341–50.

- Sugden, R. (1986). *The Economics of Rights Co-operation and Welfare*. Oxford: Blackwell.
- Sugden, R. (1989a). 'Game theory without backward induction', mimeo, University of East Anglia.
- Sugden, R. (1989b). 'Spontaneous order', *Journal of Economic Perspectives*, 3, 85–97.
- Sugden, R. (1990a). 'Convention, creativity and conflict', in Y. Varoufakis and D. Young (eds), *Conflict in Economics*. London: Harvester-Wheatsheaf.
- Sugden, R. (1990b). 'Rational bargaining', in M. Bacharach and S. Hurley (eds), *Essays in the Foundations of Decision Theory*. Oxford: Blackwell.
- Sugden, R. (1991). 'Rational choice: a survey of contributions from economics and philosophy', *Economic Journal*, 101, 751–85.
- Sugden, R. (1993). 'Thinking as a team: Towards an explanation of non-selfish behavior', *Social Philosophy and Policy*, 10, 69–89.
- Sugden, R. (2000). 'The motivating power of expectations', in J. Nida-Rümelin and W. Spöhn (eds), *Rationality, Rules and Structure*. Amsterdam: Kluwer.
- Sugden, R. (2001). 'The evolutionary turn in game theory', *Journal of Economic Methodology*, 8, 113–30.
- Svejnar, J. (1986). 'Bargaining power, fear of disagreement, and wage settlements: theory and evidence from US industry', *Econometrica*, 54, 1055–78.
- Thomson, W. (1995). 'Population monotonic allocation rules', in W. Barnett et al. (eds), *Social Choice, Welfare and Ethics: Proceedings of the Eighth International Symposium in Economic Theory and Econometrics*. Cambridge: Cambridge University Press.
- Thucydides (1953). *The History of the Peloponnesian War*. Athens: Hestia.
- Udén, L. (2001). *Methodological Individualism: Background, History and Meaning*. London and New York: Routledge.
- Udén, L. (2002). 'The changing face of methodological individualism', *Annual Review of Sociology*, 28, 479–507.
- Van Huyck, J., R. Battalio, and R. Beil (1990). 'Tacit coordination, strategic uncertainty, and coordination failure', *American Economics Review*, 80, 234–48.
- Varoufakis, Y. (1986). 'Optimisation and strikes: Microeconomic models of industrial disputes', PhD Dissertation, Department of Economics, University of Essex.
- Varoufakis, Y. (1989). 'Worker solidarity and strikes', *Australian Economic Papers*, 28, 76–92.
- Varoufakis, Y. (1990a). 'Solidarity in conflict', in Y. Varoufakis and D. Young (eds), *Conflict in Economics*. Hemel Hempstead: Harvester Wheatsheaf.
- Varoufakis, Y. (1990b). 'Conflict in equilibrium', in Y. Varoufakis and D. Young (eds), *Conflict in Economics*. Hemel Hempstead: Harvester Wheatsheaf.
- Varoufakis, Y. (1991). *Rational Conflict*. Oxford: Basil Blackwell.
- Varoufakis, Y. (1992). 'Modelling Rational Conflict', *Economie Appliquée*, XLV, 53–78.
- Varoufakis, Y. (1992/3). 'Freedom within Reason: From axioms to Marxian praxis', *Science and Society*, 56, 440–66.
- Varoufakis, Y. (1993). 'Modern and postmodern challenges to game theory', *Erkenntnis*, 38, 371–404.
- Varoufakis, Y. (1996). 'Bargaining and strikes: Towards an evolutionary framework', *Labour Economics*, 3, 385–98.
- Varoufakis, Y. (1997). 'Moral rhetoric in the face of strategic weakness: Experimental clues to an ancient puzzle', *Erkenntnis*, 46, 87–110.
- Varoufakis, Y. (1998). *Foundations of Economics: A Beginner's Companion*. London and New York: Routledge.
- Varoufakis, Y. (2002a). 'Against equality', *Science and Society*, 66, 448–72.
- Varoufakis, Y. (2002b). 'Deconstructing *homo economicus*? Reflections on an encounter between postmodernity and neoclassical economics', *Journal of Economic Methodology*, 9, 389–96.
- Varoufakis, Y. (2005). 'Rational rules of thumb in finite dynamic games: *N*-person backward induction with inconsistently aligned beliefs and full rationality', *American Journal of Applied Sciences* (Special Issue), 57–60.
- Varoufakis, Y. (2006). 'The bonds that impede: A model of the joint evolution of corruption and apathy', *Indian Journal of Economics*, 54, 84–105.
- Varoufakis, Y. (2008). 'Capitalism according to evolutionary game theory: On the impossibility of a sufficiently evolutionary model of historical change', *Science and Society*, 72, 63–94.
- Varoufakis, Y. (2009). 'Pristine equations, tainted economics and the postwar order', paper presented at the *Cold War and the Social Sciences* workshop, Columbia University, 10 April.
- Verbrugge, R. (2006). 'Nonergodic corruption dynamics (or, Why do some regions within a country become more corrupt than others?)', *Journal of Public Economic Theory*, 8, 219–45.
- Vilks, A. (1992). 'A set of axioms for neoclassical economics and the methodological status of the equilibrium concept', *Economics and Philosophy*, 8, 51–82.
- von Neumann, J. (1928). 'Zur Theorie der Gesellschaftsspiele', *Mathematische Annalen*, 100, 295–320. Reprinted in 1959 as 'On the theory of games of strategy', in Tucker A., and R. Luce (eds) *Contributions to the Theory of Games*, Vol. 4. Princeton, NJ: Princeton University Press.
- von Neumann, J. (1937). 'Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes', in K. Menger (ed.), *Ergebnisse eines mathematischen Kolloquiums, 1935–36*. Leipzig and Vienna: Franz Deuticke. Published in English in 1945 under the title 'A model of general economic equilibrium' in the *Review of Economic Studies*, 13, 1–9.
- von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.
- Weibull, J. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press. Weintraub, R. E. (2002). *How Economics Became a Mathematical Science*. Durham, NC: Duke University Press. Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.
- Wolff, R. P. (1981). 'The derivation of the minimal state', in J. Paul (ed.), *Reading Nozick*. Oxford: Basil Blackwell.
- Wood, E. M. (1989). 'Rational Choice Marxism: Is the game worth the candle?', *New Left Review*, 177, 41–88.
- Wood, E. M. (1995). *Democracy against Capitalism: Renewing Historical Materialism*. Cambridge: Cambridge University Press.

Young, H. (1993). 'An evolutionary bargaining model', *Journal of Economic Theory*, 59, 145–68.

Zeuthen, F. (1930). *Problems of Monopoly and Economic Warfare*. London: George Routledge and Sons.