*Journal of Economic Methodology 7:1, 33–57 2000*

# Rational choice, functional selection and empty black boxes

*Philip Pettit*

**Abstract** In order to vindicate rational-choice theory as a mode of explaining social patterns in general – social patterns beyond the narrow range of economic behaviour – we have to recognize the legitimacy of explaining the resilience of certain patterns of behaviour: that is, explaining, not necessarily why they emerged or have been sustained, but why they are robust and reliable. And once we allow the legitimacy of explaining resilience, then we can see how functionalist theory may also serve us well in social science; we lose the basis – the empty black box ar gument – on which the rational-choice critique of the theory has mostly been grounded.

**Keywords:** rational, functional, selection, virutal, explanation

## 1 INTRODUCTION

Those of us who have welcomed rational-choice theory as a way of doing social science have often sourced that enthusiasm in a critique of the functionalist theory that it supplanted (Elster 1979). But this dual attitude of enthusiasm and critique has proved hard, at least in my own case, to sustain. For it turns out that in order to vindicate rational-choice theory as a mode of explaining social patterns in general – social patterns beyond the narrow range of economic behaviour – we have to recognize the legitimacy of explaining what I have described as the resilience of certain patterns of behaviour (Pettit 1993, 1995). And once we allow the legitimacy of explaining resilience, then we can see how functionalist theory may also serve us well in social science; we lose the basis on which the rational-choice critique of the theory has mostly been grounded (Pettit 1996).

There is a common problem that rational choice theory and functionalist theory each have to confront. I call this the problem of the empty black box. So far as I can see, both approaches are going to fail in face of this problem or they are both going to find resources for overcoming it; they are going to sink or swim together. Drawing on earlier work, I shall argue here that the problem can be overcome in the case of rational-choice theory but that the solution offered directs us to a parallel solution in the case of functionalist theory.

My paper is in five sections. In the first I present the spectre of the empty

black box that haunts functional explanations in social science. In the second I show that there is a similar spectre that haunts rational-choice explanation in social science. In the third I show that the spectre can be dispelled in the rational-choice case by recognizing the propriety of explaining the resilience of social patterns as distinct from their emergence or continuity. And then in the fourth section I show that a similar move can be made to dispel the spectre that hangs over functionalist theory. Finally, in the fifth section, I offer some general comments on the style of explanation exemplified in each case.

## 2 THE PROBLEM FOR FUNCTIONAL EXPLANATION

Functional explanation in biological science offers the obvious model on which to think about such explanations in social science. Why do we find such and such a trait in this or that sort of organism? Why do we find beating hearts, or echolocating devices, or tit-for-tat patterns of behaviour, in this or that species or population? The answer given is that the trait serves a certain function: it circulates blood, or makes it easy to find food, or it helps individuals to achieve mutually beneficial cooperation. The very fact of serving such a function, the very fact of conferring the sort of benefit in question on its bearers, is meant to explain why the trait is found in individuals of the relevant type.[1]

Such functional explanations are tolerated in biological science, because it connects fairly obviously with the theory of natural selection. Suppose that a trait, T, is held to be functional in producing an effect, F, and that the disposition to produce F is regarded as offering an explanation for why we find T in relevant organisms. That picture of things becomes a plausible hypothesis under a paraphrase in terms of the mechanics of natural selection. The paraphrase, roughly cast, goes like this: the accidentally induced mutation whereby the gene for T appeared in the ancestors of the organisms in question gave those creatures an advantage over competitors in producing offspring, and in increasing the frequency of T in the population; it did this, in particular, so far as T-bearers manifested the effect, F. Why then do we find T in the population or the species or whatever? Well, because T produces F and because that gave T-bearers an advantage in the natural selection stakes: in short, because T is functional, so far as it produces F, because T has the function of producing F (Neander 1991a, b).

The biological model of functional explanation suggests that the aim of functional explanation in social science is to explain why certain social traits are to be found in this or that society or institution or whatever, as the biological analogue explains why certain traits are to be found in this or that species or population or whatever. And the availability of a natural selection mechanism to make sense of functional explanation in biology raises the question as to what sort of mechanism underlies functional explanation in social science. The empty black box problem is that for most functional explanations in social science there is no obvious mechanism to cite and that the

explanations, therefore, are apparently baseless (Elster 1979).

Why do we find religious rituals in various societies? Because they have the function of promoting social solidarity (Durkheim 1948). Why do we find common ideas of time and space, cause and number (Durkheim 1948; Lukes 1973: 442)? Because they serve to make mental contact and social life possible. Why do we find certain peacemaking ceremonies in this or that culture? Because they serve to change the feelings of the hostile parties towards one another (Radcliffe-Brown 1948: 238–39). Why do we find social stratification – the unequal distribution of rights and privileges – in modern societies? Because it makes it possible to fill socially indispensable but individually unattractive positions (Davis and Moore 1945).

The problem with all of these bread-and-butter examples of functional explanation is that it is not clear why the fact that the trait in question has the functional effect cited explains why the trait is found there, explains why we find the relevant religious rituals or peacemaking ceremonies or structures of social stratification. It is not clear what mechanism is supposed to operate in the black box that links the functionality of the trait with its existence or persistence. No one supposes that intentional design plays the linking role. The only mechanism that could do so appears to be a mechanism of selection akin to that which is invoked in biology; there may be other mechanisms possible in the abstract but they would not seem to fit these standard sorts of cases (Van Parijs 1981). And in most cases there is no evidence of a mechanism of selection having been at work.

There are some examples, it is true, where functional explanation in social science can be backed up by a selectional story. Some economists say that the presence of certain decision-making procedures in various firms can be explained by their being functional in promoting profits and they back up that explanation with a scenario under which the firms with such procedures, being the firms which do best in profits, are the ones that survive and prosper; they are selected for the presence and effects of those procedures in a competitive market (Alchian 1950; Nelson and Winter 1982). But it is very implausible to think that such selectional mechanisms are available for social-functional explanation in general (Pettit 1993: 155–63). The black box which functionalist thinkers apparently have to postulate is in most cases empty.

## 3 THE PROBEM FOR RATIONAL-CHOICE EXPLANATION

Rational-choice theory, as I understand it, is the attempt to use economic models of explanation in areas that go beyond what is traditionally seen as economic behaviour. In order to see that there is a problem for such theory that parallels the problem raised for functionalist theory, it is necessary to look at the assumptions which economics make about the way agents produce the behaviour it seeks to explain about the contents of the black box in the head of *Homo economicus*.

There are two sorts of assumption that economists make about the minds of the agents with whom they are concerned. First, process-centred assumptions about the way in which desires or degrees of preference issue in action. And second, content-centred assumptions about the sorts of things that the agents desire: about which things they prefer and with what intensities.

As for process-based assumptions, the first thing to notice is that economists almost universally accept the relatively weak claim that whenever people act, they do so as a result of their own desires or utility functions. They do not act on the basis of moral belief alone, for example; such belief issues in action, only if accompanied by a suitable desire. And they do not act just on the basis of perceiving what other people desire; the perception that someone desires something can lead to action only in the presence of a desire to satisfy that other person. Some thinkers toy with the possibility that agents may be capable of putting themselves under the control of something other than their own desires: for example, Mark Platts (1980) when he imagines that moral belief may motivate without the presence of desire; Amartya Sen (1982, Essay 4) when he speaks of the possibility of commitment; and Frederic Schick (1984) when he canvases the notion of sociality. But economists are probably on the side of common sense in urging that all action is mediated via the desires of the agent (Pettit 1993, ch. 1).

How do people's desires lead to action, then, according to economists? The general assumption is that desires lead to actions via beliefs about the options available, about the likely consequences of those options, and so on. More specifically, the assumption is that that they lead to actions that serve the desires well according to such beliefs; in other words, that they lead to subjectively rational actions. There are different theories as to what it is for an action or choice to serve an agent's desires well, according to his or her beliefs: about what it is for an action to be subjectively rational. But the family of theories available is exemplified by the Bayesian claim that an action is rational just in case it maximizes the agent's expected utility (Eells 1982).

So much for the assumptions that economists make about the way desires or preferences lead to action. What now of the assumptions that they make about the content of what human beings prefer or desire? The main question here is how far economists cast human beings as egocentric in their desires.

Many economists endorse what is sometimes known as non-tuism. They hold that people's desires in regard to others are not affected by their perceptions of other people's desires – utility functions are independent (Gauthier 1986: 87). Or they hold, more strongly still, that not only do people take no account of what others desire in forming their own desires in regard to others; any desires they have for what others should do, or for what should happen to others, are motivated ultimately by a self-centred desire for their own satisfaction (Gauthier 1986: 311). Economists endorse non-tuism to the extent that various economic models assume that any good I do you is, from my point of view, an externality for which ideally I would want to extract payment: an

external benefit that I would ideally want to appropriate for myself (or 'internalize') (Gauthier 1986: 87). But this seems to be a feature of particular models and not an assumption that is essentially built into the economic way of thinking. And it is a feature that affects only some of the standard results of the theories in question, not all of them (Sen 1982: 93). I am not inclined to regard it as a deep feature of economic thinking. It may have little or no presence, for example, in the application of economic thought to social life outside the market.

But even if economics does not require people to be non-tuistic, even if it allows that they may have non-instrumental desires in relation to others – perhaps desires that are affected by their perception of what the others desire – still it does generally assume that there is something egocentric about the desires on which people generally act. Economists assume that people's self-regarding desires are generally stronger than their other-regarding ones: that in this sense people are relatively self-regarding in their desires. Whenever there is a conflict between what will satisfy me or mine and what will satisfy others, the assumption is that in general I will look for the more egocentric satisfaction. I may do so through neglecting your interests in my own efforts at self-promotion, or through helping my children at the expense of yours, or through jeopardizing a common good for the sake of personal advantage, or through taking the side of my country against that of others. The possibilities are endless. What unites them is that in each case I display a strong preference for what concerns me or mine, in particular a preference that is stronger than a countervailing preference for what concerns others.[2]

The assumption that people are relatively self-regarding in their desires shows up in the fact that economists and rational-choice theorists tend only to invoke relatively self-regarding desires in their explanations and predictions. They predict that as it costs more to help others, there will be less help given to others, that as it becomes personally more difficult to contribute to a common cause – more difficult, say, to take litter to the bin – there will be a lesser level of contribution to that cause, and so on. They offer invisible-hand explanations under which we are told how some collective good is attained just on the basis of each pursuing their own advantage. And they specialize in prisoner's dilemma accounts that reveal how people come to be collectively worse off, through seeking each to get the best possible outcome for themselves.

The belief that people are relatively self-regarding shows up in other aspects of economic thought too. It may be behind the assumption of economic policy-makers and institutional designers that no proposal is plausible unless it can be shown to be 'incentive-compatible': that is, unless it can be shown that people will have self-regarding reasons for going along with what the proposal requires.[3] And it may be at the root of the Paretian or quasi-Paretian assumption of normative or welfare economics that it is uncontroversially a social benefit if things can be changed so that all preferences currently satisfied continue to be satisfied and if further preferences are satisfied as well.

This assumption is plausible if the preferences envisaged are self-regarding, for only envy would seem to provide a reason for denying that it is a good if some people can get more of what they want for themselves without others getting less. But the assumption is not at all plausible if the preferences also include other-regarding preferences, as we shall see in a moment. And so the Paretian assumption manifests a further, deeper belief that the preferences with which economics is concerned are self-regarding ones.

The Paretian assumption is not plausible – certainly not as uncontroversial as economists generally think – when other-regarding preferences are involved, for reasons to which Amartya Sen (1982, Essay 2) has directed our attention. Consider two boys, Nasty and Nice, and their preferences in regard to the distribution of two apples, Big and Small. Nasty prefers to get Big no matter who is in control of the distribution. Nice prefers to get Small if he is in control – he is other-regarding and feels he should give Big away if he is in charge – but prefers to get Big if Nasty is in control; he is only human after all. The Paretian assumption suggests, under the natural individuation of options (Pettit 1991) that it is better to have Nice control the distribution rather than Nasty. If we put Nice in control, then that satisfies Nasty, i.e. he gets Big and it satisfies Nice as well. Nice's preference for having Big if Nasty is in control does not get engaged and Nice's preference for having Small, for giving Big away, if he is in control himself is satisfied. But this is clearly crazy; it means that we are punishing Nice for being nice, in particular for having other-regarding preferences and this, while apparently attempting just to increase preference-satisfaction in an impartial manner. The lesson is that the Paretian assumption is not plausible once other-regarding preferences figure on the scene and so, if economists think that it is plausible and think indeed that it is uncontroversial, that suggests that they only have self-regarding preferences in view.

The upshot of all this, then, is that economists and rational-choice theorists present human agents as relatively self-regarding creatures who act with a view to doing as well as possible by their predominantly self-regarding desires. These desires are usually assumed to be desires for what is loosely described as economic advantage or gain: that is, roughly, for advantage or gain in the sorts of things that can be traded. But self-regarding desires, of course, may extend to other goods too and there is nothing inimical to economics in explaining patterns of behaviour by reference, say, to those non-tradable goods that consist in being well loved or well regarded (Pettit 1990; Brennan and Pettit 1993). The economic approach is tied to an assumption of relative self-regard but not to any particular view of the dimensions in which self-regard may operate.

But does the egocentric picture fit? Are human beings rational centres of predominantly self-regarding concern? It would seem not. Were human agents centres of this kind, then we would expect them to find their reasons for doing things predominantly in considerations that bear on their own advantage.[4] But this isn't our common experience, or so at least I shall argue.

Consider the sorts of considerations that weigh with us, or seem to weigh with us, in a range of common-or-garden situations. We are apparently moved in our dealings with others by considerations that bear on their merits and their attractions, that highlight what is expected of us and what fair play or friendship requires, that direct attention to the good we can achieve together or the past that we share in common, and so on through a complex variety of deliberative themes. And not only are we apparently moved in this non-egocentric way, we clearly believe of one another – and take it, indeed, to be a matter of common belief – that we are generally and reliably responsive to claims that transcend and occasionally confound the calls of self-regard. That is why we feel free to ask each other for favours, to ground our projects in the expectation that others will be faithful to their past commitments, and to seek counsel from others in the confidence that they will present us with a more or less impartial rendering of how things stand.

Suppose that people believed that they were each as self-regarding as economists appear to assume; suppose that this was a matter of common belief amongst them. In that case we would expect each of them to try to persuade others to act in a certain way by convincing them that it is in their personal interest to act in that way: this, in good part, by convincing them that they, the persuaders, will match such action appropriately, having corresponding reasons of personal advantage to do so. Under the economic supposition, there would be little room for anyone to call on anyone else in the name of any motive other than self-interest.

The economic supposition may be relevant in some areas of human exchange, most saliently in areas of market behaviour. But it clearly does not apply across the broad range of human interaction. In the normal mode of exchange, people present each other with considerations that, putatively, they recognize both as relevant and potentially persuasive. I do not call on you in the name of what is just to your personal advantage; did I do so, that could be a serious insult. I call on you in the name of your commitment to certain ideals, your membership of certain groups, your attachment to certain people. I call on you, more generally, under the assumption that like me you understand and endorse the language of loyalty and fair play, kindness and politeness, honesty and straight talking. This language often has a moral ring but the terminology and concepts involved are not confined to the traditional limits of the moral; they extend to all the terms in which our culture allows us to make sense of ourselves, to make ourselves acceptably intelligible, to each other.

Consider how best an ethnographer might seek to make sense of the ways in which people conduct their lives and affairs. An ethnographer that came to the shores of a society like ours – a society like one of the developed democracies – would earn the ridicule of professional colleagues if they failed to take notice of the rich moral and quasi-moral language in which we ordinary folk explain ourselves to ourselves and ourselves to one another: the language, indeed, in which we take our bearings as we launch ourselves in action. But if it is

essential for the understanding of how we ordinary folk behave that account is taken of that language, then this strongly suggests that economists must be mistaken – at least they must be overlooking some aspect of human life – when they assume that we are a relatively self-regarding lot.

The claim that ordinary folk are oriented towards a non-egocentric language of self-explanation and self-justification does not establish definitively, of course, that they are actually not self-regarding. We all recognize the possibilities of rationalization and deception that such a language leaves open. Still, it would surely be miraculous that that language succeeds as well as it does in defining a stable and smooth framework of expectation, if as a matter of fact people's sensibilities do not conform to its contours: if, as a matter of fact, people fall systematically short – systematically and not just occasionally short – of what it suggests may be taken for granted about them.

We are left, then, with a problem for rational-choice theory: that is, a problem for the use of economic method in explaining non-economic behaviour. We are left, in fact, with the problem of an empty black box. The mind postulated in rational-choice theory is that of a relatively self-regarding creature. But the mind that people display towards one another in most social settings, the mind that is articulated in common conceptions of how ordinary folk are moved, is saturated with concerns that dramatically transcend the boundaries of the self. So how can we invoke the workings of the economic mind to explain behaviour, when the black box at the origin of behaviour does not apparently contain an economic mind?

Rational-choice theory is in the same pickle, so it transpires, as the functionalist theory it has often aspired to supplant. Functionalist theory is apparently committed to there being a history of functional selection at the origin of the behaviours and institutions that it explains, yet there is no functional selection in evidence. Rational-choice theory seems to be committed to there being a process of self-regarding motivation and deliberation at the origin of the behaviours and institutions to which it is directed, yet the mental processes in evidence among relevant agents are not particularly self-regarding in character. [5] In each case we are invited to believe that a black-box mechanism is operating in a certain way when all the indications are that the black box is empty: or at least empty of the sort of mechanism that the theory postulates.

## 4 THE SOLUTION WITH RATIONAL-CHOICE EXPLANATION

### 4.1 The model of virtual self-regard

The problem of the empty black box that economists and rational-choice theorists face is not one of my invention (Hindess 1988). So far as rational-choice theorists have reflected on the problem, the general suggestion has

been that people are implicitly – in the sense of unconsciously – self-regarding. Gary Becker (1976: 7) comes close to endorsing this model when he writes:

> the economic approach does not assume that decision units are necessarily conscious of their own efforts to maximise or can verbalise or otherwise describe in an informative way reasons for the systematic patterns in their behaviour. Thus it is consistent with the emphasis on the subconscious in modern psychology.

But the claim that people are all unconsciously self-regarding is not particularly compelling. We all admit that people profess standards from which they often slip and that their slipping does usually relate to an awareness, perhaps a deeply suppressed awareness, of the costs of complying with the standards. We all admit, in other words, that weakness of will and self-deception are pretty commonplace phenomena. But the suggestion here is that the whole of human life is shot through with this sort of failure: that what we take to be a more or less occasional, more or less localized sort of pathology actually represents the normal, healthy state of the human organism. That is a fairly outrageous claim. Most economists and rational-choice theorists would probably be shocked to hear that the view of the human subject which they systematically deploy is about as novel and about as implausible as the picture projected in classical Freudianism.

The solution that I prefer for the problem facing rational-choice theory postulates, not that people are implicitly or unconsciously self-regarding, but that they are potentially or virtually so. Let it be granted that while actual self-regard may play a great part in market and related behaviour, it has little or no deliberative impact on the ordinary run of non-market behaviour, for example, in contexts of ordinary family or friendly interaction, in contexts of political decision, or in contexts of group behaviour. This is a worst-case scenario from the point of view of an attempt to vindicate rational choice theory. What I suggest, however, is that even under that worst-case assumption, self-regard may still have an important presence: it may be virtually if not actually there; it may be waiting in the wings, even if it is not actually on stage.

Here is how self-regard might have a virtual presence in such contexts. Suppose, first of all, that people are generally content in non-market contexts – we can restrict our attention to these – to let their actions be dictated by the cultural framing of the situation in which they find themselves, by the habits or perhaps the whims underpinned by that framing. A friend asks for a routine level of help and, in the absence of urgent business, the agent naturally complies with the request; it would be unthinkable for someone who understands what friendship means to do anything else. There is an election in progress and, the humdrum of everyday life being what it is, the agent spontaneously makes time for going to the polls; that is manifestly the thing to do, under ordinary canons of understanding, and the thing to do without thinking about it. Someone has left a telephone message asking for a return call about some

matter and the agent doesn't hesitate to ring back; even if aware that there is nothing useful they can tell the original caller, they shrink from the impoliteness in their culture of ignoring the call. In the pedestrian patterns of day-to-day life, the cultural framing of any situation will be absolutely salient to the ordinary agent and the ordinary agent will more or less routinely respond. Or so at least I am prepared to assume.

But that is only the first part of my supposition. Suppose, in the second place, that despite the hegemony of cultural framing in people's everyday deliberations and decisions, there are certain alarm bells that make them take thought to their own interests. People may proceed under more or less automatic, cultural pilot in most cases but at any point where a decision is liable to cost them dearly in self-regarding terms, the alarm bells will tend to ring and prompt them to consider personal advantage; and heeding considerations of personal advantage will lead people, generally if not invariably, to act so as to secure that advantage: they are disposed to do the relatively more self-regarding thing.

Under these suppositions, self-regard will normally have no actual presence in dictating what people do; it will not be present in deliberation and will make no impact on decision. But it will always be virtually present in deliberation, for there are alarms which are ready to ring at any point where the agent's interests get to be possibly compromised and those alarms will call up self-regard and give it a more or less controlling deliberative presence. The agent will run under cultural pilot, provided that that pilot does not carry them into terrain that is too dangerous from a self-interested point of view. Let such terrain come into view, and in most cases the agent will quickly return to manual; they will quickly begin to count the more personal losses and benefits that are at stake in the decision on hand. This reflection may not invariably lead to self-regarding action – there is such a thing as self-sacrifice, after all – but the assumption is that it will do so fairly reliably.

Under the model of virtual self-regard, most actions are performed without self-regarding consideration but that is true only so far as most actions happen to do suitably well in self-regard terms. The agent is genuinely moved by ordinary, culturally framed considerations but only so far as those considerations do not require a certain level of self-sacrifice. Let the considerations push the agent below the relevant self-regarding level of aspiration – this will vary, no doubt, from individual to individual – and the alarms bells will ring, causing the agent to rethink and probably reshape the project on hand. Otherwise put, the model gives self-regard a filtering or policing role in relation to regular, culturally framed considerations. Those considerations will hold sway in ordinary contexts but only so far as the behaviour produced in those contexts by those considerations satisfies a certain individually relative threshold of self-regard.

But is the model of virtual self-regarding control, in particular the scenario of the alarm bells, a plausible one? The question divides in two. First, is there any arrangement under which we can imagine that such alarms are put in

place? And second, if there is, can we plausibly maintain that those alarms will reliably serve to usher self-regarding deliberation into a controlling position in the generation of behaviour?

The alarms required will have to be informational; they will have to be signals that this is the sort of situation where the agent's advantage may be compromised, if habit or whim is given its head. So are there signals available in ordinary contexts that might serve to communicate this message? Clearly, there are. Consider the fact that a decision situation is non-routine; or that it is of a kind where the agent's fingers were already burned; or that it is a situation in which the agent's peers – others who might be expected to fare about as well – do generally better than the agent. Those facts are going to suggest that in such unusual or such changed circumstances the behaviour produced by the culturally framed considerations may no longer be satisfactory in self-regarding terms. Or consider the fact that while the contexts remain stable, the behaviour of the agent is changing, due to a certain drift in the effect of the culturally framed considerations or due to their being disturbed by other factors. Those facts too are going to suggest the possibility that the behaviour is not suitably satisfactory. Given that facts like these can serve as signals that the agent's personal advantage may be in especial danger, it is reasonable to assume that the alarm bells required in the model of virtual regard are going to be available.

The other question is whether it is plausible, given the availability of signals of this kind, to postulate that the signals will generally tip agents into a self-regarding sort of deliberation: a sort of deliberation that is normally sidelined in favour of fidelity to the cultural frame. This issue is wholly an empirical matter but it is an issue on which the weight of received opinion speaks unambiguously. It has been common wisdom for at least 2,000 years of thinking about politics that few are proof against temptation and few, therefore, are likely to ignore signals that their self-interest may be endangered. Human beings may be capable of reaching for the stars but, except for some romantic strands of thought, all the streams in the western tradition of thinking suggest that if there is opportunity for an individual to further their own interests, then they can generally be relied upon, sooner or later, to exploit that opportunity: all power corrupts. The main theme of the tradition is summed up in the lesson that no one can be entrusted with the ring of Gyges that Plato discusses: the ring that renders a person invisible and that makes it possible for them to serve their own interests with impunity, at whatever cost to the interests of others.

These lines of thought give support, therefore, to the picture described above. They suggest that it is very plausible to think that even where people pay no actual attention to relatively self-regarding considerations, still those considerations have a certain presence and relevance to how people behave. They are virtually present, in the sense that if the behaviour rings the alarm bells of self-interest – and there will be plenty of such bells to ring – the agent

will give heed and will tend to let self-regarding considerations play a role in shaping what is done.[6]

### 4.2 The explanatory relevance of virtual self-regard

The question which now arises, however, is how far the merely virtual presence of self-regard is supposed to legitimate the economic explanatory enterprise: the enterprise of explaining various patterns in human affairs by reference to rational self-regard.[7] If self-regarding considerations have a purely virtual presence in ordinary human deliberation – for the moment we continue to make this extreme assumption – then they are not actual causes of anything that the agents do. They may be standby causes of certain patterns of behaviour: they may be potential causes that would serve to support certain patterns, were they not supported by culturally framed deliberation. But it is not clear how anything is to be explained by reference to causes of such a would-be variety. After all, explanation is normally taken to uncover the factors operative in the production of the events and patterns to be explained; it is normally taken to require a reference to actual causal history (Lewis 1986, Essay 22).

This difficulty can be underlined by considering the explananda that economic investigation is ordinarily taken to be concerned with in the non-market area. These are, first, the emergence of certain phenomena or patterns in the past and, second, their continuation into the present and future. The explanation of the emergence of any phenomenon – say, the emergence of a norm or institution – clearly requires a reference to the factors that were operative in bringing it into existence. And the explanation of the continuation of any phenomenon, equally clearly, requires a reference to the factors that keep it there.[8] So how could a reference to virtual self-regard serve to explain anything? In other words, how can our model of the common-cum-economic mind serve to make sense of the explanatory claims of economics, in particular of the economics of non-market behaviour: of behaviour that is motored by the perception of what situations demand, under relevant cultural frames, not by considerations of self-regard?

The answer, I suggest, is that even in the unlikely event that self-regard plays no role in explaining the emergence or continuation of a pattern of behaviour – we will return to that assumption in the concluding section – still it can be of great utility in explaining a third explanandum: the resilience of that pattern of behaviour under possible disturbance or drift.

Imagine a little set-up in which a ball rolls along a straight line – this, say under Newton's laws of motion – but where there are little posts on either side that are designed to protect it from the influence of various possible but non-actualized forces that might cause it to change course; they are able to damp incoming forces and if such forces still have an effect – or if the ball is subject to random drift – they are capable of restoring the ball to its original path. The

posts on either side are virtual or standby causes of the ball's rolling on the straight line, not factors that have an actual effect. So can they serve any explanatory purpose? Well, they cannot explain the emergence or the continuation of the straight course of the rolling ball. But they can explain the fact – and, of course, it is a fact – that not only does the ball roll on a straight line in the actual set-up, it sticks to more or less that straight line under the various possible contingencies where disturbance or drift appears. They explain the fact, in other words, that the straight rolling is not something fragile, not something vulnerable to every turn of the wind, but rather a resilient pattern: a pattern that is robust under various contingencies and that can be relied upon to persist.

The resilience explained in this toy example may be a matter of independent experience, as when I discover by induction – and without understanding why – that the ball does keep to the straight line. But equally the resilience may only become salient on recognizing the explanatory power of the posts: this, in the way in which the laws that a theory explains may only become salient in the light of the explanatory theory itself. It does not matter which scenario obtains. In either case the simple fact is that despite their merely standby status, the posts serve to resolve an important matter of explanation. They explain, not why the pattern emerged at a certain time, nor why it continues across a certain range of times, but why it continues across a certain range of contingencies: why it is modally as distinct from temporally persistent.

The lesson of our little analogy should be clear. As a reference to the virtually efficacious posts explains the resilience with which the ball rolls on a straight line, so a reference to a merely virtual form of self-regard may explain the resilience with which people maintain certain patterns of behaviour. Imagine a given pattern of human behaviour whose continuation is actually explained by the cultural framing under which people view the relevant situations and by their habit of responding to that framing. Suppose that that pattern of behaviour has the modal property of being extremely robust under various contingencies: say, under the contingency that some individuals peel away and offer an example of an alternative pattern. The factors that explain its actually continuing may not explain this robustness or resilience; there may be no reason, so far as they go, why the example of mutant individuals should not display a new way of viewing the situation, for example, or should not undermine the effects of inertia. So how to explain the resilience of the pattern? Well, one possible explanation would be that as the contingencies envisaged produce a different pattern of behaviour, the alarm bells of self-interest ring and the self-regarding deliberation that they prompt leads most of the mutants and would-be mutants back towards the original pattern.

I said earlier that in all likelihood the thresholds at which people's alarm bells ring, and they begin to think in self-regarding terms, may vary from individual to individual. This means in turn that a pattern of behaviour may be very resilient in some individuals, less resilient in others and that the individual-

level explanations of resilience may not have the same force; they may not support different predictions for different individuals. But this variation, of course, need not affect aggregate-level explanation. While allowing for individual differences in self-regard thresholds, for example, we may be confident that across the population as a whole a certain general pattern of behaviour enjoys resilience in relation to a certain degree of drift or disturbance in the producing causes; people's thresholds may generally be low enough to ensure that self-regard will kick in and stabilize the pattern.

The analogy with the rolling ball serves to show how in principle the model of virtual self-regard may leave room for the economic explanation, at the level of individual or aggregate, of behaviour that is not actively generated by considerations of self-regard. But it may be useful to illustrate the lesson more concretely.

David Lewis' (1969) work on convention is often taken as a first-rate example of how economic explanation can do well in making sense of a phenomenon outside the traditional economic domain of the market. He invokes the fact that conventions often serve to resolve certain problems of coordination in explanation of such conventions; thus the convention of driving on the right (or the left) serves to resolve the coordination problem faced by drivers as they approach one another. But what is supposed to be explained by Lewis' narrative? Lewis is clearly not offering a historical story about the emergence of conventions. And, equally clearly, he is not telling a story about the factors that actually keep the conventions in place; he freely admits that people may not be aware of the coordination problem solved by conventional behaviour and may stick to that behaviour for any of a variety of reasons: reasons of inertia, perhaps, or reasons of principle or ideology that may have grown up around the convention in question.

The best clue to Lewis's explanatory intentions comes in a remark from a later article when he considers the significance of the fact that actually conventional behaviour is mostly produced by blind habit:

> An action may be rational, *and may be explained by the agent's beliefs and desires*, even though that action was done by habit, and the agent gave no thought to the beliefs or desires which were his reasons for action. If that habit ever ceased to serve the agent's desire's according to his beliefs, it would at once be overridden and corrected by conscious reasoning.
>
> (Lewis 1983: 181; my emphasis)

This remark gives support to the view that what Lewis is explaining about convention, by his own lights, is not emergence or continuance but resilience. He implies that the servicing of the agent's – as it happens, self-regarding – desires is not the actual cause of the conventional behaviour but a standby cause: a cause that would take the place of a habit that failed to produce the required behaviour in circumstances where that behaviour continued to be what self-interest required. And if the servicing of self-regard is a standby

cause of this kind, then what it is best designed to explain is the resilience, where there is resilience, of the conventional behaviour.

But it is not only the Lewis explanation of conventional behaviour that lends itself to this gloss. Can we explain American slave-holding by reference to economic interests (Fogel and Engerman 1974: 4), when slave-holders articulated their duties and conducted their business, in terms of a more or less religious ideology? Yes, to the extent that we can explain why slave-holding was a very resilient institution up to the time of the civil war; we can explain why the various mutants and emancipationists never did more than cause a temporary crisis. Can we explain the failure of people to oppose most oppressive states as a product of free-rider reasoning (North 1981: 31–32), when it is granted that they generally used other considerations to justify their acquiescence? Yes, so far as the free-riding variety of self-regarding reasoning would have been there to support non-action, to make non-action resilient, in any situation where the other, actual reasons failed to do so and alarms bells rang. Can we invoke considerations of social acceptance to explain people's abiding by certain norms, as I have tried to do elsewhere (Pettit 1990), when I freely grant that it is considerations of a much less prudential kind that keep most people faithful to such norms? Yes, we certainly can. Self-regarding considerations of social acceptance can ensure that normative fidelity is robust or resilient if they come into play whenever someone begins to deviate, or contemplate deviation, and if they serve in such cases to restore or reinforce compliance.

The upshot will be clear. We can make good sense of economic explanation, even explanation of non-market behaviour, in terms of the model of virtual self-regard whereby the economic mind is reconciled with the common mind. That model recommends itself, then, on at least two grounds. It shows that the assumptions which economists make about the human mind, in particular about human motivation, can be rendered consistent with the assumptions of commonplace, everyday thinking. And it shows that so interpreted, the assumptions motivate a promising and indeed developing programme for economic explanation: and explanation, not just in the traditional areas of market behaviour, but across the social world more generally.

## 5 THE SOLUTION WITH FUNCTIONALIST EXPLANATION

### 5.1 The model of virtual selection

But if we can have recourse in the rational-choice case to the notion of a virtual mechanism of self-regard – a mechanism that may not operate under actual circumstances but that would operate under relevant counterfactual conditions – then we can equally well help ourselves in the functionalist case to the notion of a virtual mechanism of selection. The idea would be this. Maybe there has not been any historical selection of a given type of institution for the fact that

its instances have a certain beneficial effect, but still it might be worth noting that were the type of institution in question to be in danger of disappearing – say, under disturbance or drift – then a selectional mechanism would be activated that would preserve it against that danger. The institution is not the product of actual selection, so it may be assumed – again, this is the worst-case assumption from our point of view – but it is subject to virtual selection: it would come to be selected in any of a variety of crises that put it under pressure.

The idea here is familiar from biology and extends readily to social science. Suppose we say that a certain trait is adaptive or that the gene responsible for the trait increases the inclusive fitness of the bearer in a certain environment: roughly, it increases the propensity of the bearer to replicate its genes.[9] Just saying that a trait is adaptive does not amount to saying that it has actually been selected for in a historical process. After all, a trait might be adaptive or a trait might come to be adaptive due to a change in the environment, without ever having played a role in causing its bearers to be selected. What has to be true if a trait is adaptive is that were it to be put under pressure – as it will be, of course, under ordinary evolutionary conditions – then it would cause its bearers to be selected: they would stand a better chance of replicating their genes than relevant competitors. Adaptiveness goes with being virtually, if not actually, favoured by selectional processes (cf. Bigelow and Pargetter 1987).

It is easy to imagine virtual selection at work in the social as well as the natural world. Imagine that golf clubs have emerged purely as a matter of contingency and chance: imagine that their popularity and spread has been due entirely to the brute fact that people enjoy swinging strangely designed clubs at a solid little ball and seeing how far and how accurately they can hit it. This is to suppose that golf clubs have not actually been selected for in anything like a history of competition with other institutions. Consistently with the absence of any such historical selection, however, what might well be the case is that golf clubs have certain effects, certain functional effects, such that were they to come under any of a variety of pressures, then the fact of having those effects would ensure that they survived the pressure. And if that were the case then it would be natural to say that though not the beneficiaries of actual selection, golf clubs do enjoy the favour of a virtual process of selection.

The story is not outlandish. For golf clubs do have certain effects that are functional from the point of view of members. They are expensive to run and so generally exclusive of all but the well-to-do. They are accessible from a city base. And they enable the well-to-do in any city or town to make useful business and professional contacts. What better way to establish a business or professional relationship than in the course of a relaxed round of golf? It is plausible, then, that were golf clubs to come under various pressures, were the cost of maintaining them and the cost of membership to rise, for example, still they might be expected to survive; we might not find people leaving the clubs in the numbers that such pressures would normally predict. The members of the clubs would be forced to reconsider their membership in the event of this

sort of pressure but that very act of reconsideration would make the functionality of the club visible to them and would reinforce their loyalty, not undermine it. And were some members to leave then it would become clear to them, and to others, that they lost out in doing so.

As it is reasonable to postulate that people display a virtual, if not always an actual, self-regard, so this sort of example shows that it is quite plausible to think that social life is often characterized by virtual processes of selection. Among the institutions of the society, there are many that have functional effects. And while those effects may not give us ground for thinking that the institutions were actually selected for the effects, they may well give us ground for believing that the institutions would be selected under various counterfactual conditions. The institutions are not the beneficiaries of actual selection but they do benefit from virtual selection.

## 5.2 The explanatory relevance of virtual selection

We saw in the last section that where people are possessed of virtual self-regard, then that is enough to enable us to explain the resilience of various patterns of behaviour and institutionalization by the fact that they rationally serve the self-regarding concerns of the agents involved. Thus we can explain on this basis the resilience of certain conventions, the resilience of political inaction during periods of repression, and the resilience of slave-holding in the ante-bellum south. And we can do so even in the event – the unlikely event – that the self-regard was never activated. The same explanatory lesson carries over to the present case.

The presence of a process of virtual selection enables us to explain the resilience of various behaviours and institutions by the fact that they have certain functional effects. Maybe we can't explain the historical emergence, or even the historical persistence, of golf clubs by reference to their functional effects for members; maybe there hasn't actually been any systematic selection of golf clubs for the fact of having such effects. But even in that surely unlikely case we can explain the resilience of golf clubs – as we may come to recognize that resilience in the first place – through identifying those functional effects. We can see that because of serving business and professional members in the way they do, golf clubs are fit to survive any of a variety of challenges; at least for the foreseeable future, they are here to stay.

The possibility can also be illustrated with some of the more traditional examples mentioned in the last section. Perhaps rituals emerged and survive in certain societies, or common ideas materialized and established themselves, for the most contingent of reasons. Still it may be that they are resilient by virtue of serving social solidarity or communication, since anyone inclined to give up on them would suffer an associated loss and would be drawn back in. Thus it may be possible to save the Durkheimian stories in question. And a similar analysis goes for the claim by Radcliffe-Brown, for it may well be that

peacemaking ceremonies are resilient to the extent that they mend the feelings of hostile parties for one another and that their resilience can be explained by how they function in that respect. Perhaps individuals in conflict would miss the ceremonies in the event of their having gone into decline and would seek recourse to them afresh. Or perhaps those in power in the society would see the loss associated with the decline and would insist on their restoration.

What of the example from sociology in which stratification is explained by its effect in securing high rewards for socially important but otherwise unattractive positions? This is more problematic, since everyone might notice the loss under widespread defection from stratification – assuming there is a loss – but there would seem to be a collective action predicament blocking them from individually doing anything about it. Even assuming the functionality of stratification, then, invoking that functionality will work as an explanation of the resilience of stratification only if there is some centralized agency like the government which we can expect to restore stratification under any pressures that lead to its temporary decline. Is it plausible to think that government will be disposed to do this? We need not offer a firm judgment. If it is plausible, then the functional explanation offered is a plausible account of the resilience of stratification; if it is not plausible, then the account fails.

Phenomena may be resilient so far as departures would activate rational choice calculations and tend to inhibit or reverse those initiatives. But equally, so we now see, phenomena may be resilient so far as departures would activate a concern for certain functional effects and would tend in a similar fashion to lead to inhibition or reversal.

The sort of salvation that I am holding out for functionalist theory fits well, we should notice, with the tradition of functionalism in social science. Under the salvation offered to functionalists, the explanation they seek is the sort that would identify and put aside the features that may be expected to come and go in social life and that would catalogue the more or less necessary features that the society or culture displays: those that are resilient and may be expected to survive a variety of contingencies and crises. The tradition of thinking associated with the likes of Durkheim in the last century and Parsons in this is shot through with the desire to separate out in this way the necessary from the contingent, the reliable from the ephemeral. The idea in every case is to look for the core features of a society and to distinguish them from the marginal and peripheral. Functionalist method is cast throughout the tradition as a means of providing 'a basis – albeit an assumptive basis – for sorting out "important" from unimportant social processes' (Turner and Maryanski [1979], p. 135).

It is also worth noticing, in passing, that when G.A. Cohen (1977) classifies Marx as a functionalist, the account that he gives would make good sense within the scheme of salvation on offer here. For Cohen, Marx is committed to there being' consequence laws' which assert that various institutions are supported in existence by the fact of having certain consequences. If what we mean by those institutions being supported in existence is that they are

resilient, then there need be no problem about how certain consequence laws may obtain. A consequence law will obtain precisely when the consequence is the sort of functional effect that is going to confer resilience on anything that systematically generates it.

## 6 A GENERAL PERSPECTIVE

The upshot of all this is that rational-choice and functionalist explanation have much more in common than may have been realized by proponents of either and, indeed, that the viability of each sort of explanation is secured only on a basis that is also likely to secure the viability of the other. Even if they do not enable us to explain the emergence or continuation of the behaviours and institutions they address, still both forms of theory can do well in explaining the resilience or stability of those social patterns. I turn in this final section to five general observations about the sort of explanation involved in the two theories and about their relationship.

### 6.1 First observation

The first thing I want to notice about the style of explanation in question is that it is not an esoteric form of accounting for how things are. Resilience explanation is illustrated by the staple of explanation in economics, biology and even social science: the sort that invokes the notion – itself capable of many explications – of a stable equilibrium.

Equilibrium explanation does not show how a pattern emerged or why it is present but demonstrates that the pattern is more or less inevitable, at least in a certain context, by pointing out that any ways in which it is liable to be disturbed would lead to correction. An example is R.A. Fisher's explanation of the 1:1 sex ratio in many species (Sober 1983). Fisher's idea was that if a population ever departs from equal numbers of males and females, then there will be a reproductive advantage favouring parents who overproduce the minority sex and the 1:1 ratio will tend to be restored. Such an equilibrium explanation does not offer a distinctive way of explaining things – a distinctive *explanans* – but rather a way of explaining a distinctive *explanandum.* That the sex ratio is in stable equilibrium, or that any pattern represents such an equilibrium, is a way of saying that it enjoys a particularly high degree of resilience. Being in stable equilibrium, at least for a given context, is a limit case of being resilient.

When our rational choice stories represent certain conventions, or patterns of political inaction, or forms of ownership, as resilient then they depict them, if not as stable equilibria, at least as possessed of the stability associated with many equilibria. And when our functionalist narratives display the functionality and fitness – the propensity to survive – of this or that institution, then they do much the same thing. Resilience explanations are not marginal forms

of theoretical endeavour, then; they belong firmly in the mainstream of social science. They enable us to see that certain patterns of behaviour and institutionalization are equilibrium patterns that people have learned or chanced upon and that being equilibrium patterns we can expect them to be proof against a variety of pressures.

### 6.2 Second observation

A second comment I want to make about resilience explanation, however, goes in a different direction. This is that wherever we have a convincing explanation for the resilience of a certain behaviour or institution, then we may often reasonably think that the explanation probably serves to make sense also of the survival of that pattern under past pressures.

Once we notice that a convention is such that it would survive the reflections of the self-interested, calculating agent – that it is more or less proof against the test of such reflection – then we may well conjecture that this fact will probably have played a role in the past in ensuring the survival and persistence of the convention. And once we notice that an institution like a golf club has functional effects that make it similarly proof against various threats to its existence, then we may well conjecture that this functionality may have played a role in past times in securing the continuation of the institution. The lesson is that if we find resilience explanations, we may often be directed also to factors that explain the persistence of the patterns in question: not their day-to-day continuation but their past survival in the presence of specific dangers.

### 6.3 Third observation

A third observation bears also on another variety of explanation besides the explanation of resilience. It imposes itself on us when we ask whether the resources of resilience-explanation, rational-choice or functional, also provide us with resources for explaining the non-resilience of certain patterns.

Suppose that self-interested rationality, or social functionality, are grounds for finding suitable patterns of behaviour resilient. Does it follow that patterns of behaviour that are neither rational nor functional in these senses will be non-resilient or fragile? Strictly no, it does not follow. For suppose that there are alternative patterns of behaviour in some circumstances, such that they score equally well in terms of rationality and functionality, or do not engage with rationality or functionality. It may be that one such pattern will be more resilient than others so far as it engages with a virtual controller on a par with the mechanisms we have discussed here. It may be, for example, that the moral justifiability of one such pattern of behaviour – and the unjustifiability of alternatives – will make it more resilient than the other.

We do not have to make any assumptions against the possibility of such a further source of resilience, except so far as it conflicts with the sources

discussed here. We may think that the sort of controller envisaged will sometimes overdetermine the resilience of behaviour that is already explained in rational-choice or social-functional terms. But we will have to think that the rational-choice and the social-functional mechanisms are the more powerful and that in cases of conflict that they will generally prevail. Otherwise there would seem to be no reason for singling them out and according them importance.

### 6.4 Fourth observation

But a question now arises about the relationship between rational-choice and functional explanation. While I may seem to have saved functionalist explanation by recourse to the same schema whereby rational-choice explanation is made safe, doesn't my account mean in practice that functionalist explanation is a variety of rational-choice explanation? Isn't that indeed why there seems to be little cause to consider the possibility of the two sources of resilience coming apart? Take those counterfactual conditions that put a functional behaviour or institution in danger, that thereby make the functional effects of the pattern salient to those involved and that generate fidelity to the pattern, as a result of that salience, among the agents involved. Can't we see these as nothing more or less than conditions where the alarm bells go off and where the fact that the behaviour or institution presents itself, perhaps for the first time, as satisfying in self-regarding terms and ensures that the agents will act to preserve it? And in that case, doesn't the functional explanation come across as just a special kind of rational-choice explanation? If a pattern is resilient as a result of its functional effects, so the line goes, it will be resilient for its satisfying the self-regarding concerns of relevant agents.

I am happy to admit that many functional explanations may prove to be rational-choice explanations of this kind, for they will still constitute an interesting category, by virtue of what distinguishes them from other rational-choice explanations: viz., the focus on aggregate functional effects. But as a matter of fact functional explanations need not all be instances of rational-choice explanations, at least not in any straightforward sense. They may not conflict with rational-choice explanations in the sense of postulating patterns of intentional behaviour that confound self-interest. But the factors that ensure the availability of the explanation may involve many dispositions in relevant agents over and beyond the self-regarding disposition invoked in rational-choice explanation.

Suppose that the following laws obtain in socio-political life, as I think they may do:

1 Any outrageous crime will be given publicity by the media in a society like ours;
2 The public will react with outrage to such publicity;

3 The politicians will be obliged, on pain of reducing their chance of re-election, to register and endorse that outrage in the media;
4 The only way they can effectively do this in the television sound bite, or the newspaper headline, is by calling for, or promising, harsher penalties for the sort of crime in question;
5 If penalties for any category of crime are reduced then, however beneficial the reduction in overall crime rates, there will still be some outrageous offence committed, sooner or later, in that category.

Where such laws or regularities obtain, then we can say that a regime of harsh criminal penalties is functional in placating the outrage of the public in regard to crime. Suppose that such a regime is in place, no matter for whatever reasons. Let the regime be put in danger, say because some politicians come to office who have been persuaded by criminological findings that harsh sentencing is counter-productive. The fact that no more lenient regime will serve the function of placating public outrage as effectively as the existing one means, under our assumptions, that any attempt to introduce such a regime will fail; the society will return, sooner or later, to the harsher dispensation. Such a functional explanation conforms to the schema we sketched but it is not a straightforward rational-choice explanation. Perhaps the politicians respond rationally and self-regardingly in calling for harsher penalties. But the people do not respond particularly rationally when they feel and voice outrage. The functional explanation obtains in virtue of a variety of dispositions in relevant agents, some rational, some not.

### 6.5 Fifth observation

I have argued that both functionalist explanation and rational-choice explanation have to make do, and can make do, with empty black boxes. In many cases where such explanation is offered, there is no evidence of an actual process of functional selection or an actual process of rational choice. But while it is true that no actual process of selection or choice materializes in the black box assumed, what is the case under my story is that the box contains a mechanism that is set to go into action in the event of certain conditions being satisfied.

This means that though the black boxes are empty in one sense, they are not empty in another. Let the behaviour or institution satisfy the self-interest of agents, and it will be actively chosen in the event that the alarm bells ring and the relevant agents go into self-regarding mode. Let it have certain functional effects, and it will be actively selected in the event of coming under this or that sort of challenge. The explanation of resilience that is involved in these cases does not require active process. But neither does it require any special magic.

To conclude then, if we construct an image of social science in which rational choice and functional selection play the sorts of role envisaged here,

then the emerging picture is distinctive but intuitive. We are invited to think that many patterns of human behaviour are the product of the inventive, more or less irrepressible urge of our species to try out now this, now that, variety of action and interaction. Life is a kaleidoscope in which we take the motifs of the past and play around with them, searching out new words and tales, new modes of dress and dance, new forms of religious ceremonial, new varieties of technological accommodation, and so on. We are makers and creatures of fashion. We are exemplars of *Homo ludens*: the playful human.

But while we are invited to recognize that in these respects life is an ever-changing kaleidoscope, our picture also introduces the idea that in any society, for any epoch, that there are some enduring and stable motifs. They may be ushered into being on the same spontaneous basis as all other forms of innovation and coordination but, once introduced, they stick. Within the parameters of the society or epoch in question, they serve the rational interests of individuals, or they sustain socially important functions, in a way that ensures their relative stability; it makes them more or less proof against the disturbance and drift that otherwise dominates the social kaleidoscope.

*Philip Pettit*
*Australian National University*
*pup@coombs.anu.edu.au*

## ACKNOWLEDGEMENTS

My thanks to those at the Rotterdam conference on *Fact and Fiction in Economics* for their helpful comments on an earlier draft. And my thanks also to those who offered helpful comments when the paper was presented at the Columbia University Philosophy of Science Colloquium and at a discussion group in New York University, at the Jowett Society in Oxford, and at a seminar in Northwestern University.

## NOTES

1 This reading of functional explanation in biology is not endorsed by everyone, of course (Cummins 1975). But it is the majority construal and it is the construal that is assumed in the argument against functionalist theory. Nor is our reading of functional explanation entirely unambiguous: to explain why a trait is found in a certain sort of organism, to use my terminology, may be to explain why that sort of organism has it or why the sort of organism in existence is one with that trait (Sober 1984: 147–48). I try to abstract here from that issue.

2 Notice that this conception of self-interest is consistent with the recognition of a capacity on the part of ordinary agents to identify with entities beyond themselves. See Pettit 1997, ch. 8.

3 In fairness, however, I should note that this search for incentive-compatibility could be motivated – reasonably or not – by the belief that however other-regarding most people are, policies should always be designed to be proof against more self-regarding 'knaves' (see Brennan and Buchanan 1981).

4 Some might say that under the assumption that human beings are rational centres of predominantly self-regarding concern – this, in a Bayesian sense – we ought to expect that they would be, not only self-concerned, but also calculating: we ought to expect that they would think in terms of the ledger of probabilities and utilities that figure in Bayesian decision theory. I do not go along with this. Bayesian decision theory says nothing on how agents manage to maximize expected utility; it makes no commitments on the style of deliberation that agents follow (see Pettit 1991).

5 This problem may be dismissed by some thinkers on the grounds that the literature on conditional cooperation shows how economically rational individuals may cooperate out of purely self-regarding motives (Axelrod 1984; Hardin 1982; Taylor 1987; Pettit and Sugden 1989). But that would be a mistake. This literature shows that economically rational individuals may come to behave cooperatively, not that they will come to think and talk in a cooperative way.

6 The picture of virtual self-regard may be modified by being made subject to certain boundary conditions. It might be held, for example, that the picture does not apply universally, only under certain structural arrangements: say, that it does not apply in family life, only in relations of a more public character. For related ideas see Satz and Ferejohn (1994).

7 Apart from the problem that I go on to discuss, there is an issue as to how, non-circularly, the economist is to tell the level of threat to self-interest at which an agent's alarm bells ring. I cannot discuss this problem here but would just note that it is parallel to the problem of determining an agent's aspiration level under Simon's (1978) satisficing model.

8 I ignore the requirements of potential explanation – fact-defective or law-defective explanation – as that enterprise is discussed by Robert Nozick (1974). It may be interesting to know how something might have come about or might have continued to exist under a different history, or under a different regime of laws, but the interest in question is not that which motivates ordinary economic attempts at explanation.

9 I ignore here the fact that as the fitness of a trait is normally understood, it is a function not just of how it would enable bearers to cope with certain contingencies that are taken as biologically relevant – these will not be all possible contingencies, of course – but of how probable those contingencies are.

## REFERENCES

Alchian, A.A. (1950) 'Uncertainty, Evolution and Economic Theory', *Journal of Political Economy* 58: 211–21.

Axelrod, R. (1984) *The Evolution of Cooperation*, New York: Basic Books.

Becker, G. (1976) *The Economic Approach to Human Behaviour*, Chicago: University of Chicago Press.

Bigelow, J. and Pargetter, R. (1987) 'Functions', *Journal of Philosophy* 34: 181–96.

Brennan, H.G. and Buchanan, J.M. (1981) 'The Normative Purpose of Economic "Science" Rediscovery of an Eighteenth Century Method', *International Review of Law and Economics* 1: 155–66.

Brennan, H.G. and Pettit P. (1993) 'Hands Invisible and Intangible', *Synthese* 94: 191–225.

Cohen, G.A. (1977) *Karl Marx's Theory of History*, Oxford: Oxford University Press.

Cummins, R. (1975) 'Functional Analysis', *Journal of Philosophy* 72: 741–65.

Davis, K. and Moore, W.E. (1945) 'Some Principles of Stratification', *American Sociological Review* 10: 242–47.

Dawkins, R. (1976) *The Selfish Gene*, Oxford: Oxford University Press.
Durkheim, E. (1948) *The Elementary Forms of the Religious Life*, New York: Free Press.
Elster, J. (1979) *Ulysses and the Sirens*, Cambridge: Cambridge University Press.
Fogel, R.W. and Engermann, S.L. (1974) *Time on the Cross. The Economics of American Negro Slavery*, Boston: Little Brown.
Gauthier, D. (1986) *Morals by Agreement*, Oxford: Oxford University Press.
Hardin, R. (1982) *Collective Action*, Baltimore: Johns Hopkins University Press.
Hindess, B. (1988) *Choice, Rationality and Social Theory*, London: Unwin Hyman.
Lewis, D. (1969) *Convention*, Cambridge, MA: MIT Press.
Lewis, D. (1983) *Philosophical Papers*, vol. 1, New York: Oxford University Press.
Lewis, D. (1986) *Philosophical Papers,* vol. 2, New York: Oxford University Press.
Lukes, S. (1973) *Emile Durkheim*, Harmondsworth: Penguin.
Macdonald, G and Pettit, P. (1981) *Semantics and Social Science*, London: Routledge.
Neander, K. (1991a) 'Functions as Selected Effects the Conceptual Analysis Defense', *Philosophy of Science* 58: 168–84.
Neander, K. (1991b) 'The Teleological Notion of "Function"', *Australasian Journal of Philosophy* 69: 454–68.
Nelson, R. and Winter, S. (1982) *An Evolutionary Theory of Economic Change*, Cambridge, Mass.: Harvard University Press.
North, D. (1981) *Structure and Change in Economic History*, New York: Norton.
Nozick, R. (1974) *Anarchy, State and Utopia*, New York: Basic Books.
Pettit, P. (1990) *'Virtus Normativa* Rational Choice Perspectives', *Ethics* 100: 725–55.
Pettit, P. (1991) 'Decision Theory and Folk Psychology', in Susan Hurley and Michael Bacharach (eds) *Essays in the Foundations of Decision Theory*, Oxford: Blackwell, pp. 147–75.
Pettit, P. (1993) *The Common Mind. An Essay on Psychology, Society and Politics*, New York: Oxford University Press, paperback edn with new postscript, 1996.
Pettit, P. (1995) 'The Virtual Reality of Homo Economicus', *Monist* 78: 308–29.
Pettit, P. (1996) 'Functional Explanation and Virtual Selection', *British Journal for the Philosophy of Science* 45.
Pettit, P. (1997) *Republicanism: A Theory of Freedom and Government*, Oxford: Oxford University Press.
Pettit, P. and Sugden, R. (1989) 'The Backward Induction Paradox', *Journal of Philosophy* 86: 169–82.
Platts, M. (1980) *Ways of Meaning*, London: Routledge.
Radcliffe-Brown, A.R. (1948) *The Andaman Islanders*, Glencoe, Ill: Free Press.
Satz, D. and Ferejohn, J. (1994) 'Rational Choice and Social Theory', *Journal of Philosophy* 91: 71–87.
Schick, F. (1984) *Having Reasons An Essay on Rationality and Sociality*, Princeton: Princeton University Press.
Sen, A. (1982) *Choice, Welfare and Measurement*, Oxford: Blackwell.
Simon, H. (1978) 'Rationality as Process and as Product of Thought', *American Economic Review* 68: 1–16.
Sober, E. (1983) 'Equilibrium Explanation', *Philosophical Studies* 43: 201–10.
Sober, E. (1984) *The Nature of Selection*, Cambridge, MA: MIT Press.
Taylor, M. (1987) *The Possibility of Cooperation*, Cambridge: Cambridge University Press.
Turner, J.H. and Maryanski, A. (1979) *Functionalism* , Menlo Park, CA: The Benjamin/Cummings Publishing Co.