THE VIRTUAL REALITY OF *HOMO ECONOMICUS*

The economic explanation of individual behaviour, even behaviour outside the traditional province of the market, projects a distinctively economic image on the minds of the agents involved. It suggests that, in regard to motivation and rationality, they conform to the profile of *homo economicus*. But this suggestion, by many lights, flies in the face of common sense; it conflicts with our ordinary assumptions about how we each feel and think in most situations, certainly most non-market situations, and about how that feeling and thought manifests itself in action. What, then, to conclude? That common sense is deeply in error on these matters? That, on the contrary, economics is in error—at least about non-market behaviour—and common sense sound? Or that some form of reconciliation is available between the two perspectives? This paper is an attempt to defend a conciliationist position.

The paper is in four sections. In the first section I describe the economic mind that is projected in economic explanation, whether explanation of market or non-market behaviour. In the second section I argue that this is not the mind that people manifest in most social settings and, in particular, that it is not the mind that common sense articulates. In the third section I show that nevertheless the economic mind may have a guaranteed place in or around the springs of human action; it may have a virtual presence in the generation of action, even action on which it does not actually impact. And then in the last section I show that where the economic mind has such a virtual presence, that is enough to license an important variety of economic explanation: the explanation of the resilience or robustness of certain patterns rather than the explanation of their emergence or continuance. I believe that this is the variety of explanation which is pursued generally in the economics—if you like, in the rational-choice explanation—of social or non-market behaviour.

"The Virtual Reality of *Homo Economicus*" by Philip Pettit,
The Monist, vol. 78, no. 3, pp. 308–329.

1. The thesis of the economic mind

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

The process-centred assumptions boil down to the assumption that people's actions serve their desires well, given their beliefs about such matters as the options available, the likely consequences of different options, and so on. There are different theories as to what it is for an action or choice to serve an agent's desires well, given the agent's beliefs: about what it is for an agent to be rational. Many economists work with relatively simple models but the family of theories available is usefully exemplified by Bayesian theories of rationality (Eells 1982). According to Bayesian theory, an action is rational just in case it maximises the agent's expected utility.

The Bayesian idea is that every agent has a utility function that identifies a certain degree of utility, a certain intensity of preference, for every way the world may be—every prospect—and a probability function that determines, for each option and for each prospect, the probability that the choice of that option would lead to the realisation of that prospect. An action will maximise the agent's expected utility just in case it has a higher expected utility than alternative options, where we determine the expected utility of an option as follows: We take the prospects with non-zero probability associated with the option; we multiply the utility of each prospect by the fraction representing the probability of its being realised in the event of the option's being chosen; and we add those products together.

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. In order to discuss it, we need some distinctions between different theses that each ascribe a certain egocentricity.

1. *Self-centredness.* This relatively weak claim says that people do what they do as a result of their own desires or utility functions. They do not act

on the basis of moral belief alone; 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.

2. *Weak non-tuism.* This is a stronger claim, in the sense that it presupposes the first but represents people as intuitively more egocentric still. People's desires bear on how others behave and on what happens to others, so the thesis goes, but such desires are not affected by perceptions of what those agents desire, even for themselves; people's utility functions, as it is often put, are independent of one another (Gauthier 1986, 87).

3. *Strong non-tuism.* A stronger claim again: people's desires do not extend, except instrumentally, to others. 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 desire for their own satisfaction (Gauthier 1986, 311).

4. *Self-regardingness.* A thesis that presupposes 1 but represents an alternative way of strengthening it to that represented by 2 and 3. People's non-instrumental desires may extend to others, and they may be responsive to the perceived desires of others—2 and 3 may be false—but the more that the desires bear on their own advantage, the stronger they are; in other words, people are relatively self-regarding in their desires.

Economists almost universally accept the first, self-centredness thesis. Agents who are rational in any economically recognisable sense cannot be led to action just by moral belief or the perception of what another desires or anything of the sort; such belief or perception may affect what they do but only through first affecting what they desire. 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). In any case, that is what I shall assume in what follows. There is a conflict between economics and

common sense, as I shall be arguing, but it does not arise in respect of this first thesis.

Do economists go beyond the rather uncontroversial form of self-centredness articulated in thesis 1? They certainly do so to the extent that certain versions of the axioms of consumer-choice theory go beyond minimal requirements of rationality, self-centredness included, and imply features like the downward-sloping demand curve. But that is not the issue. The question is whether economists go beyond the postulate of self-centredness in postulating any of the more egoistic theses, 2 to 4.

Many economic theories endorse weak and strong non-tuism. They do so 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 "internalise") (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 1986, p. 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.

Some may say that there is a deeper reason than the frequent use of non-tuistic assumptions for thinking that economic thinking is strongly non-tuistic in nature. The deeper reason, according to these theorists, is that in holding that agents act so as to satisfy their desires, economists assume that agents act for the sake of achieving their own desire-satisfaction: that is, for the sake of attaining a certain benefit for themselves. Anthony Downs gives countenance to this line of thought when, ironically enough, he tries to explain how economists can make sense of altruism. "There can be no simple identification of acting for one's greatest benefit with selfishness in the narrow sense because self-denying charity is often a great source of benefits to oneself. Thus our model leaves room for altruism in spite of its basic reliance upon the self-interest axiom" (Downs 1957, p. 37).

The line of thought in Downs's remark is confused. Accepting the economic theory of rationality may mean believing that people maximise expected utility but it does not mean believing that they act for their own greatest benefit. That persons maximise expected utility means that they act in the way that best serves their desires, according to their beliefs, but

not that they do so *for the sake of* maximum desire-satisfaction and, in that sense, *for the sake of* their greatest benefit. When I act on a desire to help an elderly person across the road, I act so as to satisfy that desire but I do not act for the sake of such satisfaction; I act for the sake of helping the elderly person. To think otherwise would be to confuse the sense in which I seek desire-satisfaction in an ordinary case like this and the sense in which I seek it when I relieve the longing for a cigarette by smoking or the yearning for a drink by going to the pub.

I am prepared to concede, then, that while economics postulates self-centredness in the sense of thesis 1, it does not necessarily suppose that people are non-tuistic in the senses defined in theses 2 and 3. But, to come now to thesis 4, I do think that the discipline is committed to the assumption 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 jeopardising 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.

The assumption that people are relatively self-regarding in their desires shows up in the fact that economists 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 specialise 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.

It may be said against this that I am focussing on purely contingent aspects of economic explanations: that there is no reason why economists

should not develop their explanations on the basis of other-regarding desires as well. Perhaps fewer people will put their litter in a bin that becomes more difficult to access. But, equally, fewer people will put their litter in a bin, if it comes to be generally believed that littering is not so bad after all: say if it comes to be believed, however improbably, that littering has some good environmental side-effects. Or so any economist should be prepared to admit.

This observation shows that economists can and should recognise the relevance of relatively other-regarding desires. But it does not demonstrate that they must take those desires to be potentially just as powerful as self-regarding preferences. And the explanatory practice of economists manifests the contrary belief. The working assumption behind economic explanation is that, however much people may care for others, care for a collective good, or care for some moral principle, their self-concern is likely to outweigh the effects of such care, if it comes into conflict with it. That is why it must be a miracle in the economics textbook if some aggregate or collective pattern emerges or continues when the available self-regarding reasons argue against people's doing the things that the pattern requires.

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.[1] 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—this, because 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—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—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 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-interested 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-interest but not to any particular view of the dimensions in which self-interest may operate.

2. The conflict with common sense

Does the picture fit? Are human beings rational centres of predominantly self-interested 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.[2] 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-interest. 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 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-interested as economists appear to assume; suppose that this was a matter of common belief amongst them. In that case we would expect much of the discourse that they carry on with one another to assume the shape of a bargaining exchange. 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. The normal mode under which people exchange with one another is closer to the model of a debate than the model of a bargain. It involves them in each presenting to the other considerations that, putatively, they both recognise as relevant and potentially persuasive. I do not call on you in the name of what is merely 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 in 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.

One way of underlining this observation is to consider how best an ethnographer might seek to make sense of the ways in which people conduct their lives and affairs. An ethnographer who 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.[3]

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 recognise the possibilities of rationalisation 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. The economic mind 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, if at all, can the economic mind be reconciled with the common-or-garden mind?[4]

3. The economic mind as a virtual presence

The obvious answer for would-be conciliationists is to say that whereas ordinary folk conform in most contexts to the picture of the common mind, the economic mind is still *implicitly* present in such contexts. But how to interpret this? What does it mean to say that the economic mind is implicitly present: that people are implicitly but not explicitly oriented towards the self-regarding concerns that economists privilege?

The main model of the implicit-explicit distinction is drawn from a visual analogy. It suggests that an explicit concern is something focal, something directed to the centre of a subject's field of vision, whereas an implicit concern is a concern for what lies at the edge of that field: a concern for what is peripherally rather than focally tracked. If I explicitly desire something, my desire is explicit in the sense in which I am explicitly aware of the computer screen in front of me; if I implicitly desire something, my desire is implicit in the sense in which I am—or was a moment ago—only implicitly aware of the telephone at the edge of my desk. Does this model help in explicating the idea that even if people are not always explicitly of an economic turn of mind, they are at least implicitly so?

The model certainly gives us a picture of what it might mean to say that implicitly people are economically minded. It would mean that even as people pay attention to the sorts of concerns engaged in ordinary exchanges with others, even as they keep their eyes on the needs of a friend, the job that has to be done, the requirements of fairness, they invariably conduct some peripheral scanning of what their own advantage dictates that they should do. The model does not deny the appearance of more or less other-regarding deliberation but it does debunk that appearance. It suggests that whether they are aware of it or not, those who practise other-directed deliberation indulge a more self-directed style of reflection in the shadows of the mind, on the boundaries of their attention. Gary Becker (1976, p. 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 maximize or can verbalize or otherwise describe in an informative way reasons for the systematic patterns in their behavior. Thus it is consistent with the emphasis on the subconscious in modern psychology."

But the focal-peripheral interpretation of the claim that people are implicitly self-regarding does not make the claim seem 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 what the focal-peripheral model would suggest 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 localised, sort of pathology actually represents the normal, healthy state of the human organism. That is a fairly outrageous claim. Most economists 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.

But if we reject the focal-peripheral way of reconciling the economic and the common mind, are we forced to choose between the two pictures of the human subject? Are we forced to choose between economic science and common sense? Happily, I think not. There is a second, less familiar model of the implicit-explicit distinction that is available in the literature and it promises a different, more attractive mode of reconciliation.[5]

I call this the virtual-actual model. One area where it is sometimes deployed—though not in so many words—is in explaining the sense in which I may implicitly believe that 2 times 101 is 202, even when I have never given a thought to that particular multiplication; or, to take another example, the sense in which I may implicitly believe that Europe has more than ten million inhabitants, when I have only ever thought about the population of individual countries. I implicitly believe these things in the sense that I am so disposed—specifically, I am so familiar with elementary arithmetic or with the population figures for European countries—that even the most casual reflection is sufficient to trigger the recognition that indeed 2 times 101 is 202, indeed the population of Europe is more than ten million. I virtually believe the propositions in question—virtually, not actually—but the virtuality or potentiality in question is so close to realisation that ordinary usage scarcely marks the shortfall.[6]

I propose that if we are to follow the familiar conciliationist route of describing people as economically minded, but not always in an explicit

fashion, we should try to spell out this claim by reference to the virtual-actual model, not the peripheral-focal one. I think that it is not implausible that people are virtually self-regarding in most contexts of choice, even if they are not actually so. It is generally agreed that actual self-regard plays a great part in market and related behaviour but that it does not have the same sort of presence—if it has a presence at all—in other contexts: for example, in contexts of ordinary family or friendly interaction, in contexts of political decision, or in contexts of group behaviour. What I suggest is that in such non-market contexts 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 what we might call the cultural framing of the situation in which they find themselves. 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 ring and prompt them to consider personal advantage; and heeding considerations of personal advantage leads people, generally if not invari-

ably, 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 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 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 at 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.

If the suppositions I have described were realised, then it would be fair to say that people are implicitly self-regarding: that they implicitly conform to the image of the economic mind. The reason is that under the model of virtual self-regard, no action is performed without self-regarding consideration unless it fails to ring certain alarms: that is, unless it promises to do suitably well in self-regard terms. What it is to do suitably well may vary from individual to individual, of course, depending on their expectations as to what is feasible and depending on their self-regarding aspirations: depending on how much they want for themselves, and with what intensity. But the point is that regardless of such variations, the model of virtual self-regarding control does privilege self-regard in a manner that conforms to the image of the economic mind. Another way of putting this point is to say that under the model described, an agent will generally be moved by certain considerations only if they satisfy a certain negative, self-regarding condition: only if they do not tend to lead the agent towards a certain level of self-sacrifice. Let the considerations push the agent below the relevant self-regarding level of aspiration and the alarm bells will ring, causing the agent to rethink and probably reshape the project at hand.

The position which self-regard is given under the model of virtual self-regarding control is rather like that which it enjoys under Herbert

Simon's (1978) model of satisficing as distinct from maximising behaviour. People do pretty well in self-regarding terms, even if they do not do as well as possible. And it may even be that virtual self-regarding control enables them to do as well as possible in egocentric terms, for the absence of self-regarding calculation in most decisions represents a saving in time and trouble—these are virtues emphasised by Simon—and it may also secure other benefits: it may earn a greater degree of acceptance and affection, for example, than would a pattern of relentless calculation.

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 cultural framing 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 in which 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; or that some conventional or other assurances as to the responses of others are lacking. Any such facts can serve as signals that the agent's personal advantage may be in especial danger. Indeed it is hard to imagine a situation in which the agent's interests were likely to be compromised in significant measure by culturally framed demands—compromised in a measure that the agent would not generally tolerate—without such signals being present. Certainly it is reasonable to assume that generally there will be signals available in such situations that the agent should take care: signals to the effect that this is a situation where that framing is liable to serve the agent less well than it ordinarily does.

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 two thousand 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, expect for some romantic strands of thought, all the streams in the Western tradition of thinking suggest that if there is opportunity for individuals 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 him to serve his 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 when 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 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.[7]

Under the emerging picture, then, there is a sense in which people are always at least implicitly of the self-regarding cast of mind projected by economists; if they are not actually self-regarding in their mode of deliberation, they are virtually so: if self-regard does not actually occupy the pilot's seat, it is always there in the co-pilot's, ready to assume control. The picture is a rather non-idealistic representation of human beings but it is not unnecessarily bleak. It emphasises that in the normal run, people are not calculating or self-concerned: they articulate their lives and relationships in the currency of received values and they generally conform to the requirements of those values. Where it goes non-idealistic, it does so only in the spirit of what we might call the Gyges axiom: the principle that virtue—fidelity to the demands of the cultural frame—is fragile and generally survives only under conditions in which it is not manifestly against the interests of the agent, only under conditions in which the alarm bells do not ring.

There are two further points to put to those who worry about the alleged non-idealism of our picture. First, the picture leaves open the pos-

sibility that in many cases some individuals will not heed the alarms and will stick to what the culturally framed situation requires, by criteria of common values, through the thick and the thin of self-sacrifice. And, second, the picture leaves room for the Aristotelian principle that people become virtuous, become lovers of virtuous ways, through habituation in those ways. It leaves room, not just for the possibility that some people will be relatively heedless of the alarms described, but for the possibility that such heedlessness may be facilitated in increasing measure by a regime in which the alarms only rarely ring: a regime in which things are well designed and people are free, in the silence of self-regard, to develop an attachment to doing that which by the common values of the culture is what the situation requires.

4. The economic mind as an explanatory principle

We saw in the first section that the economic mind is distinctively self-regarding and in the second that it contrasts in this respect with the common mind: the mind as articulated in common ways of thinking. The last section gave us a picture under which it seems possible to reconcile these two points of view: the points of view associated respectively with economics and common sense. The common-sense viewpoint is valid to the extent that ordinary folk manage their affairs most of the time without adverting to their own interests; they are guided in their decisions by what is required of them under the cultural framing of the situations in which they find themselves. The economic viewpoint is valid to the extent that even when this is so, even when people are not explicitly self-regarding in their deliberations, still self-regard has a virtual presence; it is there, ready to affect what people do, in the event that any of the alarm bells of self-interest ring.

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.[8] That self-regarding considerations have a virtual as distinct from an actual presence in human deliberation means that 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 sustain those patterns, did the actual causes fail. But it is not clear how anything is to be explained by reference to causes of such a would-be variety. After all, ex-

planation 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 brining it into existence. And the explanation of the continuation of any phenomenon, equally clearly, requires a reference to the factors that keep it there.[9] 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 while virtual self-regard may be of no use in explaining the emergence or continuation of any pattern of behaviour, it can be of great utility in explaining a third explanandum: the resilience of that pattern of behaviour under various shocks and disturbances.

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-actualised forces that might cause it to change course; they are able to damp incoming forces and if such forces still have an effect 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 an 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 perturbing forces appear and even have a temporary effect. 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 returning to the straight line. But equally the resilience may only become salient on recognising 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 or, more prosaically, by people's sheer inertia. 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 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—this, because of the contrast between what different individuals are doing—and the self-regarding deliberation that they prompt leads most of the mutants and would-be mutants back towards the original 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

of behaviour that is not actively generated by considerations of self-regard. But it may be useful, in conclusion, to illustrate the lesson more concretely.

David Lewis's (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—problems of a kind that can be nicely modelled with game-theory techniques—in explanation of such conventions. But what is supposed to be explained by Lewis's 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 desires according to his beliefs, it would at once be overridden and corrected by conscious reasoning" (Lewis 1983, p. 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 failing habit, so long as the behaviour remains suitable; this, in the way that Lewis says it would displace the remaining habit at the point where the behaviour becomes unsuitable. 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, p. 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. pp. 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 alarm 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 program for economic explanation: and explanation, not just in the traditional areas of market behaviour, but across the social world more generally.[10]

Philip Pettit

Research School of Social Sciences,
Australian National University, Canberra

NOTES

1. 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.

2. 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 maximise expected utility; it makes no commitments on the style of deliberation that agents follow. See Pettit 1991.

3. We may note in passing that there is nothing surprising in the fact that our ordinary encounters with one another are articulated and shaped by a non-egocentric language. We are not just bargaining creatures who take one another's beliefs and desires as given and seek out minimal terms of cooperation. We are creatures who also try to influence what we each believe and desire, under the assumption that when obstacles do not get in the way—when there is nothing we are disposed to fault about our circumstances—then we are susceptible to the same considerations in the formation of our beliefs and desires: under the assumption, equivalently, that we are sensitive to the same norms of belief and desire formation (Pettit 1993, ch. 2). Given that we pursue this enterprise, it is only to be expected that we should have evolved a language for framing culturally shared expectations.

4. This problem may be dismissed by some thinkers on the ground that the literature on conditional cooperation shows how economically rational individuals may cooperate out of purely self-regarding motives (Axelrod 1984, 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.

5. I explore a third model of this distinction in a forthcoming paper on "'Practical and Intellectual Belief" but the third model does not seem to have any application relevant to present concerns.

6. The implicitness of my belief that 2 times 101 is 202 should not be confused with the implicitness of my belief, say, that for any number described in decimal notation, you get double that number by following the sort of rule that you and I apply in computing 2 times 101: a rule, as it happens, that we would probably find it hard to articulate. The implicitness of the belief in the rule does not lend itself to modelling on the virtual-actual pattern, but rather on some other analogy such as that provided by the peripheral-focal picture, because it is clear that you and I do actually believe in the rule; we do actually believe in it to the extent that we do actually rely on it. The implicitness of my belief that 2 times 101 is 202 is the implicitness of non-actuality, the implicitness of a belief that hovers on the edge of realisation, not the implicitness of a belief that is realised in some sub-articulate fashion.

7. 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).

8. 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.

9. 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.

10. This paper represents a further development of a theme in Pettit (1993). It overlaps in some part with the text of three lectures given at the Ecole des Hautes Etudes en Sciences Sociales, Paris, and published as "Normes et Choix Rationnels," *Reseaux*, no. 62, pp. 87–112. I was greatly aided in preparing the final draft by comments received from Uskali Mäki, Raimo Tuomela and an anonymous referee.

REFERENCES

Axelrod, Robert 1984. *The Evolution of Cooperation* New York: Basic Books.

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

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

_____ and Philip Pettit 1993. "Hands Invisible and Intangible," *Synthese*, vol. 94, 1993, pp. 191–225.

Downs, Anthony 1957. *An Economic Theory of Democracy*. New York: Harper.

Eells, Ellery 1982. *Rational Decision and Causality*, Cambridge: Cambridge University Press.

Fogel, R. W. and S. L. Engermann 1974. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown.

Gauthier, David 1986. *Morals by Agreement*. Oxford: Oxford University Press.

Lewis, David 1969. *Convention*. Cambridge, MA: M.I.T. Press.

_____ 1983/86. *Philosophical Papers*, vols. 1& 2. New York: Oxford University Press.

North, Douglas 1981. *Structure and Change in Economic History*. New York: Norton.

Nozick, Robert 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Pettit, Philip and Robert Sugden 1989. "The Backward Induction Paradox," *Journal of Philosophy* vol. 86, 1989, pp. 169–82.

_____ 1990. "*Virtus Normativa*: Rational Choice Perspectives," *Ethics*, vol. 100, pp. 725–55.

_____ 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.

_____ 1993. *The Common Mind: An Essay on Psychology, Society and Politics*. New York: Oxford University Press.

Platts, Mark 1980. *Ways of Meaning*. London: Routledge.

Satz, Debra and John Ferejohn 1994. "Rational Choice and Social Theory," *Journal of Philosophy* vol. 91, pp. 71–87.

Schick, Frederic 1984. *Having Reasons: An Essay on Rationality and Sociality*. Princeton, NJ: Princeton University Press.

Sen, Amartya 1982. *Choice, Welfare and Measurement*. Oxford: Blackwell.

Simon, Herbert 1978. "Rationality as Process and as Product of Thought," *American Economic Review* vol. 68, pp. 1–16.