SCOTTISH INSTITUTE FOR RESEARCH IN ECONOMICS

SIRE DISCUSSION PAPER

SIRE-DP-2015-83

Representations and the Corruption of Goods

Martin K Jones

UNIVERSITY OF DUNDEE

www.sire.ac.uk

Representations and the Corruption of Goods

Martin K. Jones[1]

Abstract: The traditional view of the economic agent is of an individual who is self-interested, rational and perceives the world "correctly". However, there is a lot of experimental and other evidence that undermines this view of agents. It is argued that an attempt to model these agents properly requires the cognitive science idea of a *mental representation*- a mental state with content. It is shown that this idea gives economists resources to discuss critiques of economics by Sandel (2012) and Grant (2012). In particular, it allows us to think clearly about the notion of goods being "corrupted" by a change in context from a non-market to a market situation.

[1] Division of Economic Studies, School of Business, University of Dundee. DD1 4HN

1) Introduction

One of the great prejudices of modern neoclassical economics has been the belief in the universality of self-interested rational choice as a method by which one can make behavioural *and moral* judgements about how people choose. This has led to a peculiar bias against the notion that agents can perceive the same situation in different ways and come to different behavioural and moral conclusions. I will argue in this paper that the conventional theory has prevented economics from fully embracing the cognitive revolution in the behavioural sciences that may cast light on these issues. Furthermore, incorporating cognitive elements gives us resources to intelligently discuss some recent attacks on fundamental economic concepts from outside the discipline. In particular we will discuss criticisms that focus on the inability of economics to model changes in the attitudes of agents.

The aim of this paper is not to provide a defence against such attacks but to argue that we need to import concepts from cognitive science to properly discuss these criticisms. In particular, I will argue that we need to incorporate the notion of a *mental representation* into our modelling framework. A mental representation is an interpretation of the world held internally by a human being. This is an analogy used in cognitive science to representations held by information processing devices such as computers It will be demonstrated that, given the existence of our commonly- accepted ideas of how humans behave, we must accept the existence and importance of representations. I will then argue that representations are useful in untangling some of the issues in the above-mentioned critiques of economics and that they perform better than other possible methods.

Mental representations are a central part of the cognitive model of the human mind. This picture of the mind is that of the mind as a computer with inputs, data storage, processing and outputs. In this view of how the mind operates it is necessary to have entities that store data and which can be operated on by processing units. Mental representations fulfil this role within the human mind, while bytes do the same within computers (Fodor 1987). This paper advocates the use of this model of the human mind within economic reasoning.

This contrasts sharply with the "economic tradition" which lasted between the 1930s and the 1980s which tried to exclude psychological considerations from economic theory (see Bruni and Sugden 2007). The founders of marginal analysis such as Jevons and Edgworth

explicitly used psychological foundations for their ideas and tried to base their principles on what were then seen as the most modern ideas in psychology. However, Pareto in the 1930s effectively dismantled this link between economics and psychology by arguing that economics should be a science of logical action and that psychology is not necessary for economic theorising. Pareto's arguments won through and became commonplace within economics until the 1980s. One side effect of this, when the cognitive revolution emerged in the 1960s, was the ignoring of many of the concepts of cognitive science, including representations, in favour of rational choice based methods.

Despite the recent behavioural turn in economics, much of the theoretical research has focussed on creating modifications to the current theory. In ideas on social preferences, for example, the utility function tends to be modified to take account of these social preferences (e.g. Rabin 1993, Fehr & Schmidt 1999). When looking at preference consistency, modifications are made to the belief and desire functions (e.g. Tversky and Kahneman 1992). This paper will focus on issues relating to changes in attitude and will suggest that the introduction of mental representations (or just "representations" from now on) allows a much easier way of dealing with such problems than other methods, as well as bringing economics more in line with modern cognitive science.

It should be noted that throughout this paper it will be assumed that preferences may be social *or* self-interested preferences. The ideas expressed here apply to both types of preference and do not make any real distinction. It is assumed that agents will tend to behave in a self-interested or social way depending on the *reasons* that they have for acting in one way or another (in a manner to be explained). Given the wide range of experimental and theoretical papers that have been published asserting and experimentally justifying the existence of social preferences (see for example Fehr & Gachter 2000), I will not give a justification for this assumption.

2) Corruption of goods

The issue that we will focus on in this article, and which we will use as a starting point for our explanation of the role of representations, comes from the notion that goods can be *corrupted* by being handled in a different institutional setting from where they originated. This forms the basis for critiques of marketisation and the use of incentives by the philosophers Michael Sandel (2012- reviewed by Besley 2013) and Ruth Grant (2012). The

principal idea is that goods, by being bought and sold in a market or by being "bought" through an incentive, essentially change their nature and lose value (in the general, non-price sense) as a result of this[2]. This is not a novel idea since it goes back to Albert Hirsch's (Hirsch 1977) ideas on the social limits to growth.

To put this in context, we will give examples from Sandel and Grant, then look at potential explanations for the corruption of goods and ask if they hold up under scrutiny. Sandel's first example is the notion of a gift- why are gifts in Western countries usually given in kind rather than in monetary form? As Sandel points out, money is a better gift because it is fungible- it can be used by the recipient to buy whatever they want. However, a money gift is usually disliked in Western culture because it is seen as demonstrating a lack of care and attention by the giver. By converting gift-giving into a simple monetary transaction one is undermining the whole value of the gift. A more extreme example is that of friendship- one cannot buy friendship because a bought friend is simply not a friend at all.

Another example is that of paying students to learn by giving them a monetary incentive to achieve high grades. The main argument in favour of this is, of course, to improve educational outcomes. There are a variety of arguments against it. One of the main arguments is the so-called crowding out argument: the monetary extrinsic motivation crowds out any intrinsic motivation. Essentially education is no longer being desired for its intrinsic merits but for monetary outcomes. There is a large literature on this (See Bowles & Polania Reyes 2012 for a review) but, in essence, the outcome is that the nature of the good changes when making it into a marketable good. Instead of being something done for its own sake, education becomes a good that is traded for money.

A more extreme example is that of bribery. Bribery is the use of incentives in order to achieve an outcome favourable to the briber. An example would be to bribe a judge to let a person have a reduced sentence. As Grant points out, this actually satisfies most of the conventional moral strictures of modern economics- bribes are non-coercive and increase welfare for both briber and recipient (assuming that they are not caught). In strict economic terms, there is an increase in welfare. The objection comes from elsewhere: bribery undermines the norm of justice that a judge is supposed to enforce. If justice is bought and sold then its value as a fair means of punishing crime or resolving disputes is destroyed.

[2] It should be noted that both Sandel and Grant have wide ranging critiques of economic methods of which the notion of corruption of goods is only part.

An example from Sandel comes from a study by Frey and Oberholzer-Gee (1996) relating to a referendum in 1993 on whether or not to have a nuclear waste repository near to the village of Wolfenschiessen in Switzerland. In a study prior to the referendum, the villagers were given a survey on whether they would vote to accept the nuclear waste repository or not. The bulk of residents indicated that they would accept it. In a subsequent survey the villagers were asked the same question and offered an incentive of an annual monetary payment to accept the depository.

On a pure economics point of view, one would have expected the incentives to reinforce the altruistic attitude shown by the villagers but surprisingly, the villagers substantially rejected the offer. In this case it is argued, civic duty is simply corrupted by the offer of an incentive by converting it into a market transaction. The value of accepting the good because one was fulfilling one's public duty was undermined by offering the incentive.

The final example relates to the provision of an exit option. Quite often economists believe that if people disapprove of a particular arrangement for providing a good then they can express their disapproval by exiting from that market. As a result of this, if they stay in the market then they implicitly approve of any changes made. It follows that if people do tend to stay in the market then the good in question has not been corrupted as long as they have an exit option available and they have not used it. However, as Peter (2004) points out, this is to confuse choice with consent. One can have a low opinion of a particular way of distributing a good without wanting to exit from the distribution of the good.

An example of the above may be seen in the provision of health services in the UK's National Health Service. Over the past twenty years there has been an increased tendency for NHS services to be provided by private providers. This, it is claimed, reduces costs for the NHS and, since it is funded by tax-payers who also form the bulk of its users and the voting population, this would seem to be beneficial[3]. However, this has been met by hostility from many quarters in spite of this. Nevertheless, most people who have the necessary money have not tried to opt out of the system and go on to another method of providing healthcare.

This paper will accept the interpretation put on these cases by Sandel and Grant as being valid. It will be granted that these are examples of the corruption of the goods or services and the norms underlying them. Rather than debating them, we will instead focus on

[3] This is working on the assumption that these changes are cost reducing while maintaining the same level of service.

how this corruption and the reaction to it can be modelled within economics. In other words, the aim of this paper will be to find a proper analysis of the corruption of goods as a preliminary stage towards a meaningful discussion of the idea.

3) Alternative theories of corruption of goods

In this section we will examine some possible alternative ideas for the corruption of goods to see whether they are sufficient to explain the phenomenon. One set of explanations that can be eliminated immediately is a "type" based explanation derived from Harsanyi (1967, 1968a, 1968b). The type based explanation would split people into two types depending on whether they viewed goods as being market based or non-market based. However, it is hard to see how we could move on from here as the point is that individuals *change their minds* about a good and are not of a fixed type as chosen by nature. Given that types are set by nature the reason why goods can change nature in the perception of agents is highly obscure. Many of the examples given above simply would not work if agents had fixed types.

A more promising line of enquiry would seek to model corruption of goods as a change within the good itself. This was the case with Hirsch who postulated an extension of Lancaster's idea (Lancaster 1966) that goods are consumed for the sake of their characteristics rather than for being goods themselves. Under this model, there is a consumption technology that converts the consumption of goods into characteristics that the consumer wants. Utility, therefore, represents preferences over characteristics rather than goods. Hirsch postulated that this could be extended to include social norms and the environmental conditions under which they are used. In such a case, the corruption of a good would take place when these social norms were undermined or when the environment of a good is changed.

A similar, if more radical, version of this idea can be derived from the famous paper by Stigler and Becker (1977). Stigler and Becker were not concerned with the corruption of goods but rather with how to model seeming changes in tastes or preferences. Their argument was to insist that tastes should be modelled as remaining constant while all changes should be the result of changes in shadow prices. However, one can see an application of their argument to modelling corruption of goods. One way in which corruption of a good could be modelled would be as a change in taste with respect to that good. Since this involves a change in tastes,

which is unacceptable in modelling terms, it should instead be modelled as a change in a term within the utility function or, rather, a change in a term within the production function of each "commodity" within the utility function. A "commodity" in this case is actually a construct including all possible variables that could influence one's utility. This includes the original good, the alternatives to that good plus human capital goods related to it as well as other variables. The utility function itself does not change but variables within the "commodity" do change.

Stigler and Becker's own chosen variable for changes in the commodity production function, human capital, is not much use in this case as there is the possibility that a person may change their mind back to their original formulation. This would imply that acquired human capital is actually lost or forgotten. However, Stigler and Becker do allow for the possibility for other inputs influencing their commodity production function so this is not fatal for the application of their theory.

Cowen (1989) has pointed out that the refusal of Stigler and Becker to allow a given utility function to change simply pushes the change back into the commodity production function or, if that is constant, into explaining exactly why one of the variables in the commodity production function varies. Stigler and Becker provide no explanation for this change. To do so would require specifying a function to explain the change in variables as well as the commodity production function. This would provide an over-complex explanation for a change in attitudes.

A final possibility is in the use of non-standard utility functions as done by Fehr and Schmidt (1999) and others. In this case the utility of a good is explicitly influenced by the impact on other people and by social norms etc. All of these are explicit in the utility function rather than being a type of "commodity" as with Stigler and Becker. This method may seem to be an improvement over the latter approach since it does incorporate factors which they ignore and it does allow for changes in taste.

However, there are problems with this viewpoint. We can see this by looking at Fehr and Schmidt's revised utility function for two players i,j:

$$U_i(x) = x_i - \propto_i max\left|x_j - x_i, 0\right| - \beta_i max\left|x_i - x_j, 0\right|$$

Where x_i and x_j are the payoffs for agents i and j and it is assumed that $\beta_i \leq \alpha_i$. This utility function generates the utility gained by an individual in one specific situation. This varies as the payoffs increase or decrease for the individual or his opponent. However, as it stands this is not a good model for corruption of a good. For an individual to go from being inequity averse to purely payoff- oriented, for example, would (in general) require a change in the values of the parameters such that $\beta_i = \alpha_i = 0$. However, this is not explained by the model presented here- it is merely an externally imposed change in parameters.

One could also argue that any *change* in attitude will involve the addition of extra payoff values, changes in parameters and/or extra parameters to any given utility function whether self-interested or otherwise. This is because such a change in attitude requires at least two possible utility levels. This can only be achieved by shifting a parameter within the utility function to change from one utility to another. However, the addition of extra parameters that need explaining suggests that in many situations this may not be explanatorily efficient. This, of course, also ties in with the critique of Stigler and Becker offered by Cowen in that the change in Stigler and Becker's commodity production function is also unexplained. Similarly, Hirsch's adaptation of Lancaster suffers from the same problem in that the crumbling of the social norm supporting a good is likewise not explained.

This is particularly the case if, as we will explain in more depth later in the paper, the attitudes are actually discrete from each other. In other words, there may be a jump from one situation where a person is being inequity averse to another situation where that person is being self-interested. In such a situation, it would be more efficient to take account of this discreteness by simply imposing the utility numbers in the different situations rather than creating a complex function to account for the attitudes and shifts in attitudes. The latter course of action would have too many unexplained parameters and exogenous shifts in those parameters.

4) Representations, attitudes and actions

One of the more important reasons for reviewing the way in which attitudes are modelled in economics is that many of the theories in the previous section simply locate the change in attitudes as a property of outcomes. In the case of Hirsch, as well as Stigler and Becker, there is an attempt to identify changes in attitudes in changes in the properties of goods. I would argue that this tactic is hopeless since attitudes are not located in goods at all.

Attitudes are *human* mental states and are not in any sense properties of the goods themselves. In trying to incorporate differing attitudes into the goods themselves, these authors are simply not using natural categories of the world.

The modified utility approach is better since it does acknowledge that human desires and beliefs (or "tastes") do change. However changes tend to be driven by parameters that are expressed as constants. These constants have not been investigated in the sense that they do not seem to correspond to any valid and reliable psychological measure. Furthermore, there seems to be a danger of an infinite regress with parameters being explained by further unexplained parameters *ad infinitum*. If a shift in utility is explained by a shift in parameters then something is needed to explain why these parameters shifted. This requires further parameters to shift in further explanatory functions[4]. Finally, this approach seems to ignore the role of humans as active thinking beings, able to make judgements about different contexts.

In order to think seriously about what it means to take an action, how people choose one action over another and why they choose to take actions we need to analyse action from the point of view of philosophy of mind and cognitive science. The seminal paper on action in philosophy of mind is that of Davidson (1963). According to Davidson, intentional actions are *caused* by reasons, with the reason triggering the action by the human being. Reasons in turn can be split into two parts- a belief component and a desire component. So far this fits in fairly well with conventional expected utility theory with beliefs being measured by (subjective) probabilities and desires being measured by utilities[5,6].

The next stage of the argument is due to Fodor (1987) who highlighted the ubiquity of so-called folk psychology in our reasoning. Folk psychology is the tendency to attribute an action by another person to mental states in that person. In other words, if a person carries out an intentional action then we tend to believe that it is because they have a reason to do it. If this is correct then we have to accept that there is a mental state behind every action which is a psychological disposition towards a specific content. So, for example, if a person playing cricket swings his bat to hit a ball then onlookers assume that this is intentional and that he is

[4] One general methodological point here is the application of Ockham's razor: Entities must not be multiplied beyond necessity. By stipulating multiple mysterious parameters one is simply failing to explain anything.
[5] The intentional part is important here. If an action is unintentional then there is not necessarily any reason attached to it. This would be the case if someone did something accidentally.
[6] In this paper I will ignore any issues associated with measurement of preferences.

attempting to score runs. The psychological disposition here is the determination of the player to score runs, while the content is the concept of scoring runs.

Any mental state that has content is a representation and it is Fodor's main contention that any intentional causation of an action must be accompanied by an explicit representation in the mind. As Fodor points out, this links in very closely with the cognitive picture of the mind. The cognitive picture of the mind makes a close analogy between the human mind and a computer. In order for a computer to operate, it needs a representation of the data from the external world for purposes of storage and processing. A human mind, by this analogy, requires exactly the same thing. Fodor notes that this analogy is probably the only one available and it is the one that fits in best with the folk psychological picture.

Fodor tends to restrict the notion of representations to beliefs on the grounds that representations are about content and that content tends to be about what is the case. However Smith (1987) has pointed out that one need not restrict representations so tightly since content could be about how the world *should* be rather than just about how it is. In other words, one could be motivated to an action by one's goals, which would involve rearranging the world to fit what one wants to happen. In order to do this, one would need a representation of one's goals and desires.

This all suggests that intentional actions are caused by reasons which split up into beliefs and desires as Davidson claims and this can be modelled by expected utility theory. However, we can go further and posit that our reasons for actions are bound up in mental representations that include both belief elements that show the world as it is and desire elements that represent the world as one wants it to be. We can go further with the belief side of representations. Rips (1994) points out that the reasoning process amounts to a causal interaction between belief- based representations. Insofar as the contents of representations consist of mental symbols or a "language of thought" then we can see this as being a logical process resulting from the syntax of the symbols in representations.

This idea of a reason-based choice has been picked up by Shafir et al (1993) who do not explicitly claim that choices have their roots in representations but simply that reason-based choice is a different way of modelling choice from value- based choice theories such as expected utility. The claim here is that the two can be modelled together and there is no need to distinguish between them. However some of the things that Shafir et. al claim do make sense here. For example, the use of representations and reasoning does allow us to deal better

with framing effects since the representation simply represents the external world and the reasons for a choice can be seen in how it is interpreted. A framing effect can simply be seen as a "guided" interpretation of the external world in a representation.

A human being can potentially entertain a wide variety of representations, including many different representations of the same situation. We will assume that, given a choice set comprising a given number of possible actions, we can amalgamate the representations of each feasible action together into one consistent representation for the choice set. We will make two assumptions to support this:

Common Interpretation of Content: Within the constraints of bounded rationality, all relevant external events relating to the choice set must be interpreted into representational content in the same way.

Minimal Rationality: Within the constraints of bounded rationality, all reasons given for each action within the choice set must be logically consistent with the reasons given for every other action within the choice set.

Both of these follow from the general psychological desire for consistency and to avoid explicit contradictions in one's reasoning (Rips 1994). Note that both of them explicitly take account of bounded rationality. In other words humans lack processing power and knowledge and so may not be able to *recognise* some inconsistencies.

We can give examples for each of these assumptions in order to motivate them. Common interpretation of content simply states that one cannot use different interpretative frameworks within a representation. So, for example, it would be illegitimate to refuse to eat pork because one is a Muslim but also refuse to eat beef because one is a Hindu. These are distinct religious practices that preclude a person belonging to both of them. They also prevent people using both when reasoning about a given decision since they insist on different ways of interpreting the world.

Minimal rationality refers to the rationality of the *content* of the representations rather than the interpretive framework. A simple example is in the choice of fruit. One can give one's reason for choosing an apple as being because one likes apples. However, it would not then be consistent to reject choosing a pear because one did not like fruit. It follows that a representation over a choice set will be internally consistent and share a common interpretation of the external world. From now on, when we refer to "representations" we will

refer to representations over whole choice sets that abide by these consistency criteria and if we want to refer to content associated with individual actions we will simply refer to them as "reasons".

A given choice set may have a variety of different possible representations giving different reasons for taking one action or another. This is quite consistent with a wide variety of psychological research. Simply by taking notice of different cues or by believing certain things to be more important than others, one can come to very different conclusions about a given situation. We will assume, therefore, that humans can entertain a wide variety of different representations over a given choice set. Such ideas have been discussed in economics by Amartya Sen in papers on description (Sen 1980) and positional objectivity (Sen 1993).

Since we will be looking at applying this notion of representations to game theory we will need to define the equivalent of a choice set in a game. We will call this a *base game*. A base game is a game structure G=<P,S,Q> where P is the set of n players, S is an n-tuple of pure strategy sets (one for each player) and Q is a set of outcomes. A base game has no explicitly defined payoffs but only outcomes specifying the event that occurs. The players' preferences over the outcomes are not defined within the base game. Instead, representations provide the preferences for the base game, creating a new, conventional, game every time a representation is applied.

5) The Importance of Representations

Having established the necessity for accepting representations, one is entitled to ask why this is important. Why has this never been deemed to be important before? In order to establish importance, we will need to tie representations more tightly to the traditional theory of choice. Representations have as content the reasons for making a choice that include both belief and desire elements. It follows that, for a given representation over a choice set, one would expect these belief and desire elements to be measured by probabilities and utilities respectively. Each action in the choice set would result in a number of possible outcomes, each weighted by a subjective probability of that outcome occurring. The outcomes would be in turn measured in terms of the decision maker's preference over other possible outcomes by a utility index.

It follows that a representation can be interpreted as an allocation of expected utilities to a choice set or base game. This may seem trivial since expected utilities are allocated to choice sets and games all the time. However, it should be noticed that, unlike conventional decision theory, there is more than one possible representation for each choice set or base game. It may be possible for the same choice set to have completely different allocations of utilities according to the representation selected. It should be noted, as will be discussed later, that the actual judging of utilities does not happen merely because one has a particular representation but rather through the process of selection of representations.

It can be seen now why the notion of representations is important and why it has not been used before. Until the last thirty years in economics it was generally assumed that individual agents interpreted the world in the same, correct, way i.e. that the world was "description invariant" and could be described in just one way. This generally meant that there was no scope for individual interpretation. In addition, the assumption of self interest in economics has generally stopped any consideration of changes in attitude by the agent. These two factors has reduced the scope of using representations to insignificance. It is only with the acceptance that individuals may vary in both their interpretations of the world and in their attitudes that representations become important.

6) The Selection of Representations

One question that remains is how one representation is selected over another. How is it that one comes to allocate one set of utilities to outcomes in a choice set or base game rather than another? Call the representation that is actually used to allocate the utilities in a choice set or base game the *active representation*. One issue that must be tackled is the principle of minimal rationality. This applies *within* a representation but not necessarily *across* representations even if these representations apply to the same choice set. In order to maintain consistency, one would not hold two contradictory representations at one time. It follows that a reasonable human being will, in general, not have two active representations at one time and if they are aware of more than one such representation then they must make a choice between them.

Another issue is how many representations will be available for selection as the active representation. At first it might be thought that this is an impossible task to decide. There are, potentially, an infinite number of ways in which one can interpret a particular situation and

one might be tempted to say that we are only constrained by our imagination. However, there are a variety of practical constraints. Firstly, we are boundedly rational with limited time for formulating alternative representations in our mind. Our attention time is rationed and we have limited processing time for some of the more complex ideas. Secondly, we tend to look to other people for many of our ideas and we tend to adopt representations that have been formulated elsewhere. These tend to be limited in number because they have to be easily transmitted between individuals (see Sperber 1996 for a detailed analysis of this). Thirdly, the minimal rationality constraint imposes minimal standards of consistency that must be adhered to by any representation. Finally, representations have to make sense in terms of one's own experience and knowledge. Given this, as was mentioned earlier, one would expect representations to be discrete from each other. Changing a variable in a representation arbitrarily may make the whole representation incoherent either internally or with the external world. Within these constraints it is likely that the number of representations available for choosing will be comparatively small.

The question, then, is how this selection is made. The proposal that I wish to make in this paper is that this selection is the result of an autonomous choice made by the agent. In other words, representations are chosen in much the same way that one chooses goods. While there are some restrictions on how we can model these choices, I believe that there is no intrinsic problem with this method. Human beings are self-aware and are able to analyse their own beliefs and desires to see whether they fit into a given situation. In the process of this analysis they need to make a judgement as to which representation best fits the situation and to *choose* between those representations that are available.

I would contend that this is not an unusual way of thinking. Whenever we go into a new situation we are always looking for clues as to how we *should* behave and also what sort of situation it is. If we visit a foreign country we try to find out and understand the social structures, institutions and mores of that country. We then try to find the best way of behaving and do our best to conform to it. An extreme version of this is shown in the Asche experiments (Asche 1940) where it was demonstrated that conformity could force subjects in experiments to make basic misjudgements about line lengths. This happened because, apart from one real subject in each experimental session, all the other subjects were confederates who were instructed to make the same obvious mistakes. Even if we do not endorse the full conformist interpretation, it must be obvious that subjects in Asch's experiments must have

been taking into account what the other (supposed) subjects were doing and analysing the situation to make their own decision.

In order to understand this we need tools to explain what is going on. One useful concept is that of a *mental action* (Proust 2001, Geach 1957, O'Brien & Soteriou 2009)[7]. A mental action is an intentional action by the mind that has as its goal another mental process. Examples of this are easy to find. A person tries to remember where his keys are. A student tries to concentrate in a lecture. An electrician tries to work out the course of electric wiring in a wall. In all these cases nothing physical is happening but in each case the processes have goals that are intentional and are not the result of unconscious thought. If, for example, a person succeeds in remembering where his keys are then the action has been successful but not otherwise.

Given that mental actions resemble ordinary physical actions, they share the attributes of physical actions. Since they are intentional then they are caused by reasons comprising beliefs and desires and, given our use of the expected utility model, these can be measured using probabilities and utilities respectively. The mental action that we are interested in is that of the mind fitting a particular representation to the situation in which the agent finds themselves (Proust 2009). This "checking procedure" inside the mind is not solely related to correspondence between the beliefs in representations and the external world. Since accuracy is valued by most people (see Williams 2002) this would also be incorporated in one's desires. It can also involve wishful thinking whereby one's desires influence how one perceives the world[8]. One could, therefore, have expected utilities attached to the action of checking representations which are higher depending on how closely they are perceived to correspond to the external world and, in the case of wishful thinking, how much one would *like* them to correspond to the external world.

Precisely how this checking mechanism works or can be conceptualised is under dispute. However, for our purposes, it is obvious that comparisons can be made between the different checking actions to see which representation corresponds closest to the external world and then this can be *chosen* as the active representation. What we have, therefore, is a two-stage decision process. In the first stage a representation is chosen in the mind and then, *given* the representation, a choice is made from the choice set or base game. The mental

[7] I fact the notion of a mental action as described here was first implicitly described by Locke (1689)
[8] One could argue that if one's representation involves normative elements then a desire element cannot be avoided.

actions will have a representation corresponding to them which gives reasons for thinking that one representation is a better fit to the world than the others and which implicitly assigns expected utilities to these outcomes.

One result of this unified decision process is that one can model the whole procedure as a two- stage decision tree or, with more than one player, an extensive form game. The utility payoffs at the end of the tree will be the aggregated payoffs both from choices in the choice sets and also from the choice between representations. It should be noted that this is a purely measurement procedure. The expected utility values in the choice set are assigned *given* the particular representation that is chosen so the fact that representations may contradict each other does not prevent utilities being assigned or being comparable across representations. All that is required is that the utilities are measured on the same scale.

The dual character of these utilities deserves to be emphasised because they will prove to be useful in the examples at the end of the paper. The "fit" of the representation to the external world is not trivial in utility terms. To take an extreme example: suppose a person is working all day earning money as a typist. They are faced with a decision as to whether carry on working in the job or not. They have two representations: one is realistic and results in a decision to carry on working as this is the only way in which the typist can support himself. The second representation is fantastical in that it assumes a fairy is going to endow the typist with a lot of money so that he doesn't need to work ever again. While the utility from the money acquired in the second representation may be enormous, it would be swamped by the expected utility gained from the comparative "fit" of the first representation and the typist's valuation of accuracy[9].

7) Representations as moves in a game

One obvious objection to this framework is that when this is conceptualised as a game then one's opponent cannot see which move has been made as it occurs inside the opponent's head. However there are three possible responses to this objection. The first response is that hidden moves are easily modelled in game theory through the use of non-singular information sets. The fact that an agent does not know how another agent has moved is easily modelled in an extensive form game. It should be noted, however, that it is assumed that all

[9] This does not preclude people from choosing the other way, although one might categorise those who do so as "fantasists".

possible representations chosen by the other agent are known even if one does not know the specific representation that was chosen.

The second response is that, for various reasons, the representation chosen will be "obvious". To take an extreme example, a person is unlikely to have a representation that will result in him wanting to drinking poison. In other examples, "obviousness" may be culturally specific. In China, for example, it is generally assumed that Chinese people are able to use chopsticks. In other cases people may consistently have a psychological bias against certain things. For example, people tend to find furry animals to be more "cute" than (say) reptiles and may be more likely to pet them.

The third response derives from Frank's (1988) work on emotions. According to Frank, emotions act as signalling devices because they are difficult (although not impossible) to fake. If a representation causes emotional arousal in an agent then the other agent may be able to detect that emotion and distinguish which representation their opponent has chosen. Emotion is costly for the person involved and, being difficult to fake, acts as a commitment device that makes the signal reliable. One would expect this method to be used quite often in personal interactions.

Another issue relating to representations as game moves is that of mixed strategies. As we have stated above, the principle of minimal rationality states that individuals cannot accommodate contradictions within a representation. However, this is part of a general psychological finding that agents do not allow explicit contradictions within their reasoning. As a result of this, an agent would not be willing to hold two contradictory representations at the same time. However, with the traditional interpretation of mixed strategies this is precisely what does happen. Given equal expected utilities for representations then one will randomise over the two representations even if they have elements in them that contradict each other[10].

However, this need not be a major problem in games and mixed equilibria can still be used. This is because there are alternative interpretations of the mixed equilibrium concept that can be used instead. One alternative interpretation was put forward by Aumann (1987) where he interpreted a mixed strategies equilibrium as being a state where players' subjective probabilities were over their opponents' chances of playing one pure strategy (or

[10] Again, it should be noted here that utilities are simply a measure of preference and are assigned conditional on one representation or another. The fact that there may be a contradiction between representations does not prevent one from measuring utilities.

representation) over another rather than randomising one's own strategy. Another alternative is to follow evolutionary game theory where a mixed strategy is interpreted as the proportions of a population who are playing a given pure strategy (or representation).

Finally, the use of representations automatically implies the use of *extensive form games* rather than normal form games. This is because there is a natural time difference between the choosing of a representation and the subsequent choosing of an option from the basic game. This is necessary because the choice of a representation informs the interpretation of the situation that results in preferences over options in the choice set. Representations must be chosen first so that options can be chosen with coherent preferences[11]. It follows from this that we will have a preference for using extensive form refinements when looking at equilibrium in these games.

8) Representations, corruption and consent

Having outlined how representations can be modelled in a game we will need to give examples of how this may be done in more concrete situations. Fortunately, as has been described, representations can be modelled in much the same way as moves in a game, if one allows for the facts that they are largely hidden from opponents and that one cannot use the traditional view of mixed strategies.

Furthermore, we need to link our discussion of representations to the corruption of goods. As we have seen, we have rejected the idea that this can be modelled as a modification of utility functions. Ultimately, whether a good has certain values is not intrinsic to the good itself or to a particular event. Rather, any kind of valuation is a human activity, done as a result of humans deciding that one good or event has value rather than another. Any corruption of a good therefore is better modelled as a *decision* by an agent that this good does not have the same value as it had before.

In discussing the modelling of representations we will model three examples derived from the discussion in section 2. The examples given will be simple in form but they will highlight some of the issues mentioned previously in this paper. The first case is that of consent (Peter 2004) where individuals find themselves in a situation where the context has

[11] In the examples later on in this paper, it will sometimes be assumed that agents already have a fixed representation or that a move by one's opponent rigidly determines one's own representation. This does not detract from the point made here.

changed while options have not. The example given was of the NHS in the United Kingdom and the change from public to private providers of healthcare. We will assume that an exit option exists but that a person who disapproves of the change in context may still not use is spite of his disapproval. The second case is that of the Swiss nuclear waste repository where individuals approve of having a nuclear waste repository without incentives but reject it when they have it. The final case relates to bribery and whether a person will accept a bribe or not.

Case 1 is modelled as follows: the government decides on what sort of service (e.g. private or public) is provided for healthcare. There are three possible options (labelled a,b- two treatments and e- an exit option) that the second player, an individual consuming the country's health service, can choose.

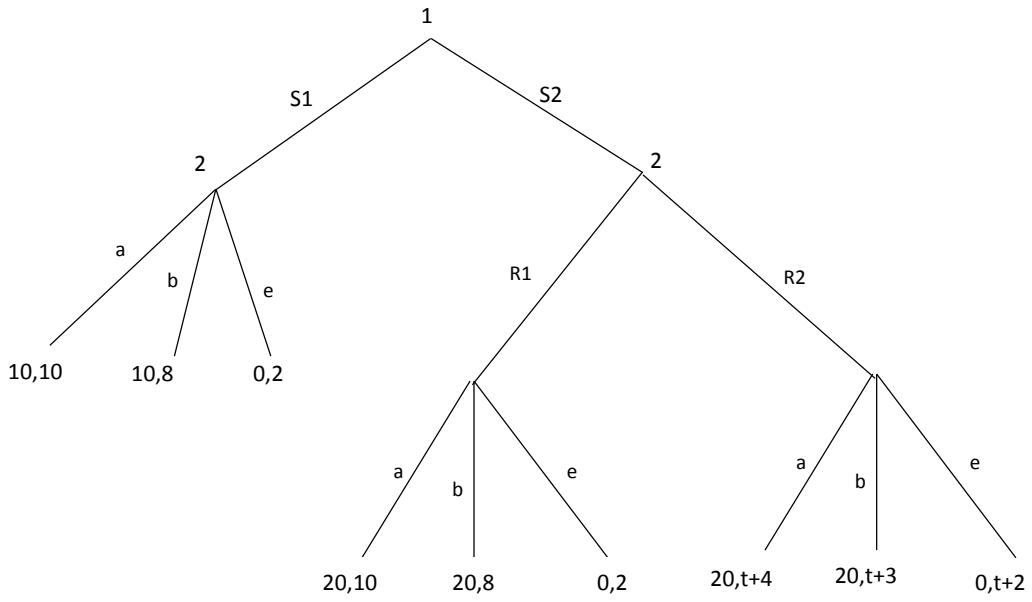

Figure 1: Healthcare game

In the diagram above we can see that S1 and S2 represent public and private provision of the service respectively. S1 is the "status quo" option in that it is assumed that the service has been public for a long time and so the second player's reaction to this style of service has settled down on one active representation. There are two possible representations: R1 and R2. R1 is the active representation that is held when public service is continued while R2 is a possible alternative representation that holds when private service is imposed. When private service is imposed then the second player has to make a choice between the two representations and then, conditional on each representation, a choice between the goods provided.

It is assumed that the government approves of the change in service and so prefers people to have private service rather than public service but doesn't like it when people exit from the service altogether. The individual's utilities are controlled by the representation they have of the situation. We will assume that individuals always prefer treatment a to b and both of these to the exit option.

Under representation R2, player 2 has an addition made to his utility payout of t. This represents the "fit" of the representation to the external world. We will normalise the corresponding value for R1 to zero. If it is judged that this representation is a good fit then t will be high. If $t > 6$ then a play of (S2,R2,a) becomes the subgame perfect equilibrium of the game. If $t < 6$ then (S2,R1,a) is subgame perfect. In both cases, whether the service remains in the public sector or not, the type of treatment chosen remains the same and from the point of view of physical actions there is no difference.

However, this should not be taken as approval of the system under which treatments are provided. If $t > 6$ under R2 then the amount of utility attached to the *treatments* has actually gone down i.e. the change in system for providing the treatments has "corrupted" them. The reason for choosing R2 is because of its perceived fit to the situation (with utility of t) not because the goods available look better. However, the individual has not chosen the exit option since it still looks like a poor alternative within the game. This, as was stated by Peter (2004), does not imply consent. Indeed, if a survey was taken then considerable dissent would probably be expressed. This means that "voice" is a necessary part of institutional design and we cannot rely on exit options alone to indicate dissent.

It should also be noted that from the point of view of the outside observer, this game would seem to be very simple. If we ignore the change in supplier then the same treatments are being produced under public or private provision. One may see this simply as a straight choice between two treatments and an exit option. We can also include the government's choice of two types of service but this does not allow us to see the thought processes inside an individual's head. This can only be done with representations.

Another aspect of this is that, once the individual is in equilibrium, he is in equilibrium in both physical actions *and in representations*. The individual is not free to simply change his interpretation of the world at whim because to do so would be to decrease his expected utility. Naturally, the individual could acquire another, third, representation but this would change the nature of the game by adding another mental action. Nevertheless, this

new game could be solved in a similar way and a representation would become fixed in the new equilibrium. Allowing for an agent to interpret a situation in a model does not lead to theoretical anarchy but instead explains the stability of these interpretations in the form of equilibrium representations.

Case 2 revolves around the notion that the Swiss state was searching for a nuclear waste repository in the village of Wolfenschiessen. In the model, the investigators either offer or do not offer money for the villagers to accept the repository. The status quo situation is that the villagers are not offered money. In this case they are assumed to subscribe to their original representation (labelled R1) and immediately decide whether to accept (A) or decline (D) the waste depository. If the villagers are offered money then they make a choice between the original representation R1 or a new one, R2. When they have chosen their representation then they accept or decline the repository.

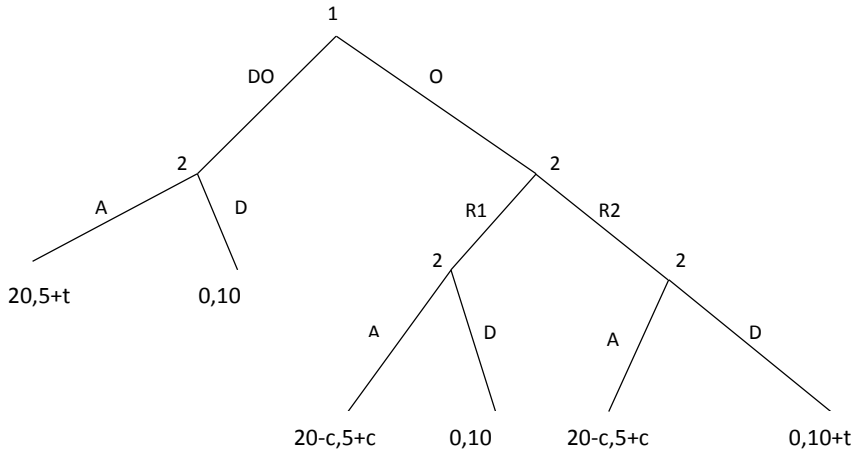

Figure 2: Nuclear Waste Game

There are two constants “c” and “t” in the game payoffs above. “c” represents the payment to the villagers, while “t” represents the utility change as a result of offering money rather than relying on public spiritedness. It should be noted here that the expected utility derived from the “fit” of the representations is not assumed to have much impact in this case. This would imply that the two representations are seen as equally plausible as a good fit for the situation. The main effect of the different representations is to change the balance of utility payoffs between A and D.

It follows that the main interest in this model is to decide under what circumstances individuals will choose D over A and R2 over R1. Under subgame perfection and assuming that money is offered, it can easily be seen that, assuming D is chosen under R2 and A under R1 then t>c-5 where c>5. If money is not offered then t>5 for A to be chosen. If the state wishes for the repository to be accepted then they will choose not to offer money. Of course, depending on the values of t and c, other outcomes are possible. However, such a set-up allows us to analyse the situation in detail.

In a similar manner to case 1, to the outside observer, the decision looks very simple- the villagers simply have to choose between accepting the waste depository and rejecting it but the use of representations shows that there is actually a rich underlying decision process. Also, in the same way as with case 1, the representations are visible to the villager making the decision but is not visible to the state so that the state cannot "see" whether one representation is chosen or another.

The final case focuses on the issue of bribery. One can assume in this case that there is a judge who has been offered a bribe in order to carry out a particular judicial decision. We will model this as a choice by a particular judge as to whether they should see the bribe as a commercial transaction (and choose R1) or as a moral issue (in which case they will follow R2). It should be noted that we are not assuming absolute morality here. Given enough money the judge will be won over or will show weakness of will[12]. We will also assume, as with the second example, that the "fit" of the representations does not affect the outcome in this example.

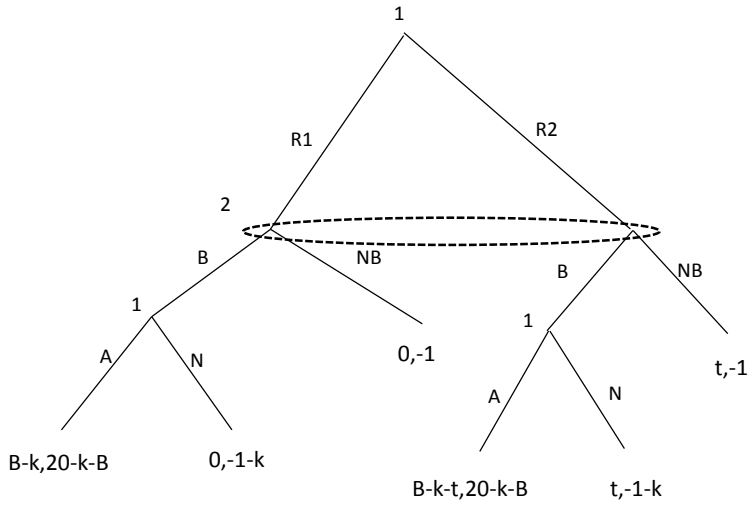

Figure 3: Bribery Game

[12] One way to demonstrate an absolute aversion to corruption would be to have the t utility set at infinity.

The choice between R1 and R2 is hidden from the briber and so there is an information set covering the starting nodes of the next decision. This is the decision as to whether to bribe the judge or not. If it is decided not to bribe the judge then the briber gets a payout of -1, while the judge gets a payout of 0 if they are corrupt but t if not corrupt. If a decision is made to bribe then the judge has to decide whether to accept the bribe or not. If the judge accepts the bribe then he gets a payout of B which is transferred directly from the briber's payout of 20. There is an expected loss k from them being caught which applies to both payouts if the judge accepts and only to the briber if the judge refuses. Finally, under representation R2 there is a utility value t that reduces the utility value of accepting a bribe if offered and boosts his utility if it is refused.

Obviously this model is highly complex and will have several different outcomes depending on the values of k, t and B. Since k is not important in the current analysis, we will fix its value at 3. We will likewise fix the value of B at 5 (so that the bribe is sufficiently large to exceed the expected costs of being caught) and see what happens when we vary t. It can be shown that the briber has a probability p=13/16 where the expected values of bribing and not bribing are equal where p is the briber's subjective probability of the judge deciding to be corrupt or not corrupt.

At a comparatively high value of t such as t=6 there is one perfect Bayesian equilibrium where the judge is not corrupt (i.e, chooses R2) and will not accept a bribe. As a result, no bribe is offered with the probability of not offering the bribe at p>13/16 . At t=0 there is no difference between the two representations- the judge treats bribes purely as market transactions under both representations. As a result, the judge's choices rely purely on the utility, the expected costs of getting caught and the size of the bribe. As a result there are two perfect Bayesian equilibria, one under each representation. In both cases the bribe is offered and the judge accepts the bribe.

The final example is when t=2 where there are three perfect Bayesian equilibria. At p<13/16 the judge is corrupt (choosing R1), is bribed and accept the bribe. Another is a mixed strategies perfect Bayesian equilibrium where player 2 is uncertain about player 1's representation, placing a probability of p=13/16 of the judge choosing R1. Player 2 bribes the judge and if the judge is not corrupt then the bribe is rejected but is accepted otherwise. The final equilibrium is where p>13/16 so the judge is not corrupt and does not accept the bribe, with the result that a bribe is not offered. It is interesting that even at low levels of t, it is still

possible for the judge to decide not to be corrupt. As can be seen, even with this simple model, it is perfectly possible for complex behaviour to emerge.

9) Discussion and Conclusion

This paper has argued that the notion of corruption of goods reveals a weakness in economics theory that needs to incorporate the notion of a mental representation into discussion of decision making. The alternative ideas to this involve modified utility functions or the use of types. However, the use of modified utility functions involves the use of mysterious psychological parameters that have little grounding in psychological reality and change for unknown reasons. Meanwhile the use of types ignores the element of choice involved in assessing a situation.

I would argue that the use of representations has many positive points in that it incorporates the latest research from the cognitive sciences. This means that advances within the cognitive sciences in terms of ethics, conceptual learning, framing, decision theory and so on can be far more easily incorporated within economics. By accepting the cognitive model of the mind as a computer (and aligning with the commonly held folk theory of psychology), economists can analyse and solve many problems that before were simply ignored or remained as puzzles.

The big advantage in using representations is that it is incorporated into game theory comparatively easily. There is no need for additional definitions of utility or new equilibrium concepts. The examples analysed above used ordinary ideas from game theory such as subgame perfect equilibrium and perfect Bayesian equilibrium. There are some restrictions on the modelling: one must use extensive form games rather than normal form games and the interpretation of mixed strategies are narrower than before. However, this does not detract from the overwhelming advantages of using representations.

The example that we have used in this paper is that of corruption of goods and it can be seen that this is modelled quite successfully through the use of representations. By using ordinary game theory one can see exactly how goods come to be corrupted. Essentially they are corrupted because they are *perceived* to be corrupted. A change of situation, whether deliberate or otherwise, undermines some of the assumptions that the agent has about the value of the good. If this undermining is seen as important i.e. if a representation with the

good's value undermined is seen as a better "fit" or if the choices under the representation are better, this representation will succeed over the others.

This also speaks to the related issue of consent. A person may choose a representation of a situation which is a good "fit". However, this does not necessarily mean that the individual likes the choices under this representation more than under another representation where the choices are allocated higher utilities. The (expected) utility from the "fit" is equally important in determining which representation is adopted. This means that a representation may be chosen even though the choices themselves are dominated by the choices under another representation. Choices therefore are an inadequate method of determining people's consent to particular choice situations.

Representations also provide a tool that can be used in modelling situations where interpretation matters. To date in economics, there has been little understanding of situations where people change their choices when they are presented in different ways. While some progress has been made in terms of negative and positive framing (c.f. prospect theory Tversky & Kahneman 1992) there has been little movement beyond that. Representations are naturally well suited to modelling interpretation and framing precisely because the *content* of representations is not constrained by the necessity to be self-interested or fully rational. Representations can include normative as well as descriptive content and so can, potentially, model a wide range of behaviour.

The models above also provide an answer as to why people tend to widely share particular beliefs and desires. These beliefs and desires, incorporated in representations, are in equilibrium with the physical choices. As a result they are fixed and remain comparatively stable. The introduction of interpretation and representations does not make modelling unmanageably heterogeneous but, surprisingly, enforces a considerable degree of uniformity. Furthermore, this explains the close links between institutions and behaviour. Given that institutions allow or disallow certain types of choice and reward, this influences the representations chosen by individuals and creates a uniformity of behaviour within that institution.

Finally, I would argue that representations are a *natural* way of looking at the world. The folk psychological view is named precisely because that is the view of the world held by ordinary human beings. It is perfectly ordinary for humans to attribute reasons to other

humans when they take actions and to do that, as has been argued here, is to assume that they have representations.

Bibliography

Asch S.E. (1940) “Studies in the principles of judgments and attitudes:II Determination of judgments by group and by ego standards” Journal of Social Psychology 12 433-465

Aumann R. (1987) “Correlated Equilibrium as an Expression of Bayesian Rationality” Econometrica 55 1 1-18

Besley T. (2013) “What’s the Good of the Market? An Essay on Michael Sandel’s *What Money Can’t Buy*” Journal of Economic Literature 51 2 478-495

Bowles S. & Polania Reyes S. (2012) “Economic Incentives and Social Preferences: Substitutes or Complements?” Journal of Economic Literature 50 2 368-425

Bruni L. & Sugden R. (2007) “The road not taken: How psychology was removed from economics and how it might be brought back” Economic Journal 117 p. 146-173

Cowen T. (1989) “Are all Tastes Constant and Identical?” Journal of Economic Behaviour and Organization 11 127-135

Davidson D. (1963) “Actions, Reasons and Causes” Journal of Philosophy 60 23 685-700

Fehr E. & Schmidt K.M. (1999) “A Theory of Fairness, Competition and Cooperation” Quarterly Journal of Economics 114 3 817-868

Fodor J. (1989) Psychosemantics: The problem of meaning in the mind Explorations in cognitive science MIT Press Cambridge MA

Frank R. (1988) “Passions within Reason: The Strategic Role of Emotions” W.W. Norton & Company New York

Frey, B.S. & Oberholzer-Gee F. (1997) “The Cost of Price Incentives: An empirical analysis of crowding out” American Economic Review 87, 4 p. 746-55

Frey, B.S., Oberholzer-Gee F.& Eichenberger (1996) “The Old Lady Visits your Backyard: A Tale of Morals and Markets” Journal of Political Economy 104, 6, p.1297-1313

Geach P.(1957) “Mental Acts: Their Content and their Objects” London Routledge & Kegan Paul.

Gneezy U. & Rustichini (2000) “A fine is a price” Journal of Legal Studies 29, p. 1-17

Grant R. (2012) “Strings Attached: Untangling the Ethics of Incentives” Russell Sage Foundation and Princeton University Press, Princeton, New Jersey.

Harsanyi J. (1967-1968) “Games with Incomplete Information played by “Bayesian” Players” Parts I-III Management Science 14, 159-182, 320-334, 486-502

Hirsch A.(1977) “Social Limits to Growth” Routledge & Kegan Paul London

Lancaster K. (1966) “A New Approach to Consumer Theory” The Journal of Political Economy 74 p. 132-57

Locke J. (1998) “An Essay Concerning Human Understanding” (first published 1689) Wordsworth editions Limited Hertfordshire

O’Brien L. & Soteriou M. (eds) (2009) “Mental Actions” Oxford University Press Oxford

Peter (2004) “Choice, Consent and the Legitimacy of Market Transactions” Economics and Philosophy 20 1 1-18

Proust J. (2009) “Is there a Sense of Agency for Thought?” in O’Brien L. & Soteriou M. (eds) (2009) “Mental Actions” Oxford University Press Oxford

Proust J. (2001) “A Plea for Mental Acts” Synthese 129 105-128

Rabin M. (1993) “Incorporating Fairness into Game theory and Economics” American Economic Review 83 1281-1302

Rips L. (1994) “The Psychology of Proof: Deductive Reasoning in Human Thinking” MIT Press Cambridge MA

Sandel M. (2012) “What Money Can’t Buy: The Moral Limits of Markets” Penguin books London

Sen A (1993) “Positional Objectivity” Philosophy and Public Affairs 22 2 126-145

Sen A. (1980) “Description as Choice” Oxford Economic Papers 32 3 353-369

Shafir E., Simonson I & A. Tversky (1993) “Reason-Based Choice” Cognition 49 11-36

Smith M. (1987) “The Humean Theory of Motivation” Mind 96 381 36-61

Sperber D. (1996) “Explaining Culture: A Naturalistic Approach” Blackwell, Oxford

Stigler G. & Becker G. (1977) “De Gustibus Non Est Disputandum” American Economic Review 67(2) p. 139-47

Tversky A. & Kahneman D. (1992) “Advances in Prospect Theory: Cumulative Representation of Uncertainty” Journal of Risk and Uncertainty 5 297-323

Williams B. (2002) “Truth and Truthfulness” Princeton University Press, Princeton, New Jersey