Triangulation across the lab, the scanner and the field

The case of social preferences

Jaakko Kuorikoski[1]
jaakko.kuorikoski@helsinki.fi
TINT – Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences
Social and Moral Philosophy
P.O. Box 24, 00014 University of Helsinki

Caterina Marchionni
caterina.marchionni@helsinki.fi
TINT – Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences
Social and Moral Philosophy
P.O. Box 24, 00014 University of Helsinki

Published in *European Journal for Philosophy of Science* 6(3): 361-376

Abstract. This paper deals with the evidential value of neuroeconomic experiments for the triangulation of economically relevant phenomena. We examine the case of social preferences, which involves bringing together evidence from behavioural experiments, neuroeconomic experiments, and observational studies from other social sciences. We present an account of triangulation and identify the conditions under which neuroeconomic evidence is diverse in the way required for successful triangulation to occur. We also show that the successful triangulation of phenomena does not necessarily afford additional confirmation to general theories about those phenomena.

[1] Corresponding author

1. Introduction

Neuroeconomics is an interdisciplinary field positioned at the intersection of economics, psychology and neuroscience. It combines theories and methods from its parent disciplines in an attempt to provide neuroscientific foundations for theories of choice behaviour, thus increasing the realisticness of microeconomic theory -- or so its advocates claim.[2] In particular, methods of data generation originally developed in neuroscience such as functional magnetic resonance imaging, positron emission tomography and transcranial direct-current stimulation are used, in combination with behavioural experimental designs, to produce evidence that could be relevant to economic claims and hypotheses. Given that the theories, methods and data span multiple levels and disciplines, neuroeconomics could be considered a blueprint for the interdisciplinary integration of experimental and observational data on the one hand, and biological, psychological and social theories on the other. Research on social preferences, to which recent neuroeconomic experiments aim to contribute, is a case in point.

The activity of bringing evidence produced by diverse methods to bear on the same hypothesis, claim or result is often referred to as *methodological triangulation*. Its main benefit is to increase confidence in the hypothesis in which multiple independent lines of inquiry converge. However, triangulation faces two major challenges that pull in opposite directions: the different types of evidence should be independent, but at the same time they should be comparable (cf. Stegenga 2009). These challenges are particularly salient in the application of triangulation to neuroeconomics. First, although the evidence generated in neuroeconomic experiments is different from traditional economic evidence in that it includes neurobiological data, it is not at all obvious that it is independent in the sense required for triangulation. Most, but not necessarily all, neuroeconomic experiments involve the scanning of subjects' brain activity while they play some standard economic game, the behavioural results of which are generally already well established. An account of what relevant independence amounts to is needed. However, whether such an account can be provided in the first place has been questioned (Hudson 2013; Schupbach 2015; Stegenga 2012). Second, the aim in many typical neuroeconomic experiments is to identify which areas of the brain are

[2] It is now quite customary to distinguish between two versions of neuroeconomics: the use of economic concepts in modelling the way neural systems work and the use of neuroscientific methods to explain behaviour related to economic choice – although the boundary between the two is becoming increasingly vague. Our main concern is with the latter.

activated when subjects carry out particular decision tasks, but why should this information be relevant to economic hypotheses? Economics is supposedly about the efficient allocation of scarce resources, not about activity in the brain. For triangulation to work it is not sufficient for the evidence to refer to the same hypothesis, it also has to be commensurable with economic evidence obtained from behavioural experiments (both in the laboratory and in the field) and in observational studies.

In what follows we investigate the extent to which these challenges can be met, and whether neuroeconomic evidence is potentially useful in the triangulation of economic hypotheses. We identify the conditions under which neuroeconomic experiments satisfy the requirements for successful triangulation – although increased confidence in the causes of observed pro-social behaviour in the laboratory is not sufficient in itself to confirm general theories of human cooperation. To this end we analyse the case of social preferences, in other words the tendency of individuals to care not only about their own material payoff or wellbeing, but also about the payoffs and wellbeing of others (Fehr and Krajbich 2014). The current debate concerns (a) whether such *preferences* exist and drive cooperative behaviour (b) in the laboratory and (c) in the wild. Although we take this case to be representative of the kind of issues raised in neuroeconomics more generally, three sets of considerations explain our focus.[3] First, multiple sources of evidence are utilized. Second, it involves several disciplines including economics, neuroscience and anthropology. Third, social preferences and theories of cooperation in general have also attracted attention in the philosophical literature (e.g. Woodward 2009, Guala 2012).

Two general lessons emerge from our analysis of this case study. First, when feasible, the integration of diverse evidence for the triangulation of phenomena is epistemically beneficial. Nevertheless, independence is not guaranteed by the fact that diverse data are generated by a different method or within a different field of study. The challenge is to understand when data produced using different methods are in fact independent in the right way. We argue that for triangulation to increase confidence in a phenomenon, the data need to be produced in causally independent processes. Whether or not two processes are causally independent

[3] This role of neuroeconomic experiments is not limited to establishing the existence of social preferences. Take, for example, *loss aversion*, a putatively fixed feature of the psychology of decision-making taken to explain many market-level phenomena in the field (Camerer 2000). Additional evidence has been gathered using brain-imaging techniques to support the claim that loss aversion is a stable property of subjective valuation. The rationale, as in the case of social preferences, is that if differences in the activation of reward-related areas are observed that match the observed asymmetric choice behaviour, then this allegedly constitutes independent evidence for the "reality" of loss aversion (see e.g. Rick 2011).

depends on the claim for which they are evidence. Second, establishing the reality of a phenomenon in the laboratory via triangulation does not necessarily warrant extra confirmation to any general theory about that phenomenon. This is important because the extrapolation of results concerning laboratory phenomenon to the wild requires theoretical backing, and a broad theoretical understanding is obviously a key reason for pursuing the interdisciplinary integration of knowledge.

The rest of the paper is organised as follows. We review arguments in favour or against the relevance of neuroscientific findings and evidence for economic hypotheses (Section 2). Next, we present our own complementary account based on the idea of methodological triangulation (Section 3). After introducing the debate on social preferences (Section 4), we examine the evidential contribution of neuroeconomic experiments (Section 5) and field experiments (section 6). In the final section, we discuss the value of diverse evidence with respect to triangulating on a phenomenon and confirming a general theory.

2. The contested relevance of neuroeconomics

The field of neuroeconomics has received a fair amount of attention both in academia and in the world outside, but it has also been the object of intense controversy. As neuroeconomist Paul Glimcher (2011) notes:

> Interdisciplinary research and controversy are nothing new in a university setting, but the birth of neuroeconomics—an attempted fusion of neuroscience, psychology, and economics of decision making—has proved unusually acrimonious. Neuroeconomics is an effort to cross a sacrosanct border in the intellectual world. (Glimcher 2011: xi)

The ideal of interdisciplinary integration collides with entrenched barriers of disciplinary identity and autonomy. Some see neuroeconomics simply as a manifestation of the arrogant scientism of economists: economics takes evidence from other disciplines seriously only when these disciplines have the appearance of hard science (such as neuroscience). What, exactly, is the relevance of the study of the brain to the study of the efficient allocation of scarce resources?

Economists Farouk Gul and Wolfgang Pesendorfer (2008) offer one of the earliest and best-known critiques of neuroeconomics. They claim that neuroeconomic data and findings are irrelevant to economics, which is strictly about observable choices and therefore only data

about choices constitute relevant evidence for economic theories. Although many commentators have argued against Gul and Pesendorfer's position, the extent to which and how neuroeconomic evidence is relevant to economics remains unclear (see e.g. Fumagalli 2013).

It has been suggested that neuroeconomics contributes to the construction of mechanistic explanations of (economic) decision-making (e.g. Craver and Alexandrova 2008; Kuorikoski and Ylikoski 2010). The idea of mechanistic integration of knowledge is in line with the stated aims of advocates of neuroeconomics, namely to provide a framework for choice behaviour in terms of neural mechanisms. It is also widely acknowledged that the standardised experimental paradigms borrowed from behavioural economics have led to some interesting neuroscientific findings concerning such mechanisms.

The idea of mechanistic integration provides a template for the production of a fuller, more comprehensive understanding of a phenomenon. Our interest here, however, is in whether neuroeconomics has *direct* evidential value for economic hypotheses. Hence, our project is closer to Clarke's (2013), who is also concerned with the relevance of neuroeconomic *evidence*. His proposal is that neuroeconomic evidence confirms economic hypotheses if and when it identifies additional variables that discriminate between competing (either psychological or behavioural) hypotheses. However, whereas Clarke's (2013) aim is to identify the conditions under which it makes sense to say that psychological (including neurobiological) evidence is *economic* evidence, in other words relevant to the confirmation of competing economic *theories*, in our view this is only part of the story. The other part builds on Woodward's suggestion that neural evidence 'provide[s] an *alternative means of triangulation on underlying causes*, such as subject's motives and preferences' (2009: 197).[4] The underlying intuition is simple: given that neuroeconomic experiments constitute a different way of experimentally 'detecting' motives and preferences, they could contribute independent supporting evidence about subjects' motives. Woodward (2009) illustrates his point with a couple of detailed examples but does not offer an account of when and why neural evidence is useful for triangulation. This is what we do next.

[4] According to Woodward (2009), neural evidence can also help in identifying the conditions under which behaviour is robust to changes. It is important to keep the two notions of robustness separate. In the case of triangulation it is a question of *detection robustness*, whereas in this case it is phenomenon robustness, in other words the robustness of the phenomenon itself. A non-robust phenomenon can still be detected robustly.

3. Triangulation, independence and comparability

Although the use of mixed methods is becoming increasingly widespread, methodological triangulation still has a somewhat dubious reputation in the social sciences (Hammersley 2008, Blaikie 1991). Philosophers of science have also raised doubts about its epistemic value (e.g. Hudson, 2013; Stegenga 2009, 2012). The word *triangulation* is considered to be a misleading metaphor and, as methodology, is seen as unduly presupposing the commensurability of the ontologies of different methods of evidence-generation. Gul and Pesendorfer's argument reviewed above could be considered a special case of the incompatible-ontologies argument: neuroscientific findings cannot be relevant evidence for economic hypotheses because there is no meaningful way in which they could be said to concern the same kinds of things. Economics is about observable choice whereas neuroeconomics evidence is not. Underlying psychological causes of choice behaviour would be relevant to economics only if the latter were *not* taken exclusively to concern the further consequences of consistent choice behaviour. We cannot contribute to the debate about the adequacy of strict revealed-preference interpretations of economic preferences here. Our account of the possible triangulating relevance of neuroscience to economics simply presupposes that, at least for some explanatory and predictive purposes, a revealed-preference interpretation is not an adequate interpretation of the role of preferences in all economic models – regardless of what the official self-understanding of economics may be (for arguments, see Hausman 2012).

At the same time, for triangulation to work, diverse evidence should be suitably independent. Not only does there appear to be a trade-off between evidence that is comparable and evidence that is independent, however, it is also unclear what kind of independence is required (Stegenga 2009, 2012; Hudson, 2013, Schupbach, forthcoming). The question of independence is especially pertinent in the case of neuroeconomic experiments, most of which simply repeat well-known behavioural experimental paradigms with well-known behavioural results. Under what conditions can such experiments be taken to provide new, *independent* evidence?

Elsewhere we have developed an account of triangulation that meets these challenges (Kuorikoski and Marchionni 2016). Our account is based on two central ideas. The first of these concerns the scope of triangulation and addresses the challenge of comparability (Stegenga 2009, 2012). In line with Bogen and Woodward (1988), we distinguish two types of evidential reasoning: data-to-phenomenon inferences and phenomenon-to-theory

inferences. Phenomena here mean robust and (at least partially) replicable results, effects or regularities. Scientific theories explain, and are tested by, phenomena. However, phenomena tend not to be directly observable and their existence must be inferred from data. The two types of inference are different: phenomenon-to-theory inferences concern the broadly logical relationship between a theory and a description of the phenomenon, whereas data-to-phenomenon inferences concern the reliability of the causal processes by which the data are generated. Hence, the methodologies employed to increase the validity of inferences from data to phenomena are different from those required to bring a variety of evidence to bear on a general theory.

This view of triangulation has two implications. First, when different data are used as evidence for the existence of a phenomenon or construct, there need not be any tight conceptual 'mapping' such as deduction or reduction between the ontologies of the phenomenon and the data. Triangulation only requires preferably controlled co-variation between the different kinds of evidence, together with plausible background theories concerning the causal processes between the phenomenon and the data. Hence, the concern that neural and behavioural data presume incommensurable ontologies does not arise. The second implication is that, although evidence about the existence of the phenomenon discriminates between theories that posit its existence and those that do not, triangulation does not necessarily provide an extra confirmational boost to the theories. As we argue in section 7, the consequence of this is that even successfully triangulating evidence concerning an experimental phenomenon does not automatically have any implications concerning the external validity of the results.

The second central idea concerns the kind of independence that successful triangulation requires. Triangulation is a form of robustness reasoning, in other words the use of independent means of determination for the purpose of controlling for errors and biases (Wimsatt 1981). Different methods triangulate only if their characteristic errors and biases are independent, which is possible if the methods are based on different causal processes.[5] If such methods produce congruent results, the results are more likely to be attributable to the phenomenon of interest rather than being artefacts of the errors and biases of particular processes. If errors are causally independent, then whether or not one method produces the correct result is independent (in a probabilistic sense) of whether or not the other produces the right result, conditional on the actual value of the hypothesis (i.e. whether or not the phenomenon exists). If this condition is satisfied, the methods are also confirmationally

[5] Schupbach (2015) calls this *reliability independence*.

independent: the confirmation provided by one is independent of the confirmation provided by the other (Fitelson 2011). This confirmational-independence condition provides the normative epistemic foundation supporting the value of triangulation, and identifies the *kind* of diversity of data needed to strengthen the reliability of the inferences from data to phenomena.

4. The debate on social preferences

We now apply this account of triangulation to identify the conditions under which neuroeconomic experiments, behavioural experiments and field observations triangulate on the existence of social preferences. The general point is that whether or not two pieces of evidence are suitably independent depends on the causal processes that produce the data, as well as on the specific hypothesis for which the data are taken to constitute the evidence.

A cursory reading of rational choice theory might give the impression that it is somewhat puzzling to suppose that people care consistently and significantly about fairness and the wellbeing of others.[6] At first sight one might also wonder how evolution could have produced organisms with concerns for others not directly related to them. Yet, our everyday experience demonstrates otherwise. Societies are not built solely on material sanctions and incentives; hence our daily social cooperation has to rely on other mechanisms as well. Laboratory experiments have also shown that people do not behave in accordance with the rational fulfilment of purely self-regarding (monetary) preferences.

As a common explanation of both sustained cooperation in the wild and the robust pro-social behavioural phenomena observed in the laboratory, it has been suggested that people have *other-regarding preferences,* or *social preferences*, understood as more or less *stable other-regarding motivational mechanisms* (H_1). Accordingly, social preferences include a diverse set of evaluations of possible outcomes that go beyond the payoff to the agent itself, such as positive outcomes of other agents/players (altruism), negative outcomes of others (punishment), outcomes for specific groups, the intentions of others, expectations of guilt, the properties of the distribution of payoffs in general (fairness), and aggregate social welfare. It should be noted that social preferences understood as relatively simple motivational states are not economic preferences in the sense of total (all-things-considered) comparative evaluations (Hausman 2012). This is important, because experimental results on social preferences in this

[6] Here and in the rest of the paper rational-choice theory is intended to encompass the assumption of self-regarding preferences.

thinner sense and economic preferences as all-things-considered judgments cannot be expected, by default, to serve the same explanatory and theoretical purpose. We believe that defining social preferences as simple motivational states is in line with what many prominent neuroeconomists say about them. It also makes them straightforwardly empirical phenomena in contrast to the more theoretical construct of preferences as all-things-considered judgments.

Social preferences as simple motivational states cannot be directly observed: their existence is typically inferred from data obtained from behavioural laboratory experiments. In such experiments, a set of subjects sits in a 'laboratory' (usually a computer class) interacting with one another via a computer. In the case of a Dictator Game (DG henceforth), the first player (proposer) simply decides whether and if so how much to give to the other player (responder), and rational-choice theory predicts that the proposer gives nothing and keeps all of the endowment. An Ultimatum Game (UG) gives the responder the option of denying the offer, in which case neither player gets any money. In this case rational-choice theory would predict an equilibrium solution such that the first player transfers the minimum amount of the endowment and the second player accepts any positive offer the first player makes. The first player in a Trust Game (TG) chooses a share to give to the other player, which is then multiplied by a set factor and the other player can then choose whether or not to give something back to the first player. Here the equilibrium solution is again that the first mover contributes nothing. Finally, in a Public Goods Game (PGG) the players contribute to a common pool of money, which is then multiplied by a set factor and distributed equally. Given the possibility of free riding, the rational prediction would again be that no one contributes. Yet, people do give and contribute. The mean offer in the DG is around 20 per cent (Camerer 2003: 57), but this consistently tends to taper off if the subjects continue to play more rounds. In the UG, not only do the proposers typically offer a substantial amount (in industrial countries almost 50%), the responders also consistently reject offers below 20 per cent (on average) of the initial endowment. The refusal of the second player to accept low offers is also decidedly irrational, in that he or she is forfeiting the possibility of financial gain for no apparent material gain. These games are usually one-shot and anonymous so as to eliminate self-interested reputation building as a possible confounder.

However, the existence of social preferences is not the only explanation advanced for the behavioural results obtained in the laboratory. 'The mistaken-framing hypothesis' (H_2) holds that the laboratory subjects are simply mistaken about the game. Anonymous one-shot UGs or DGs are such alien social situations that the subjects import familiar heuristics and

consequent patterns of play from real life, in which such social encounters tend to be repeated and not anonymous. Moreover, in repeated situations such as these apparently altruistic one-off offers and rejections are rational self-regarding strategies in the long run.
More specifically, Burton-Chellew and West (2013: 216) claim that the existence of pro-social preferences has not been tested for directly, and has only been postulated as a post-hoc explanation of observed results. They found in their experiment that people behaved no differently in the PGG when they had no idea whatsoever that they were actually playing with other people. Moreover, if the players received *more* information about the payoffs of the other players their willingness to co-operate actually decreased. Both results are clearly contrary to the social-preference hypothesis, which presupposes that knowledge of other players' outcomes motivates co-operative behaviour. Burton-Chellew and West's interpretation is that systematic patterns of behaviour observed in game-theoretic experiments are not attributable to other-regarding motivational mechanisms, but derive from simple and sub-optimal learning strategies aimed at selfishly maximizing the subjects' own monetary outcomes in the fundamentally alien laboratory environment (see Camerer 2013, however).

Finally, it has been suggested that the main drivers of behaviour are not the monetary outcomes per se, but the desire to conform to appropriate social norms that the subjects read into the experimental situation. It is not so much the payoffs, or the consequent payoff–derived social preferences, but rather the desire to avoid playing the game somehow 'incorrectly' that explains the behaviour (cf. Zizzo 2010, Gigerenzer and Gigerenzer 2005). We call this the 'please-the-investigator' hypothesis (H_3).

Hence, at least the following explanations of cooperative behaviour observed in the laboratory can be given (see also Woodward 2009): (H_1) the social-preference hypothesis; (H_2) the mistaken-framing hypothesis; and (H_3) the please-the-investigator hypothesis. Hypotheses H_2 and H_3 in effect claim that the behavioural results are not attributable to social preferences but are artefacts of the experimental set up. Several neuroeconomic studies purportedly provide triangulating evidence for the social-preference hypothesis. It is to these studies that we now turn.

5. Social preferences in the scanner

The rationale behind the study of social preferences via neuroeconomic experiments is simple: if measured or induced activation is observed in areas known to be related to reward, motivation or affect correlates with pro-social behaviour, then it is hypothesised that the

observed behaviour is intrinsically motivated by other-regarding considerations. Therefore, the detection of such activation is taken to be evidence for the existence of genuine social preferences. The evidential role of neuroeconomics rests largely on 'reverse inference' from neural activity in a particular brain region to cognitive function (Poldrack 2006). This inference is taken to be especially compelling when the level of observed pro-sociality can be associated with the level of the detected 'reward' or 'affect'. On the other hand, it is more problematic when the region in question is involved in several kinds of cognitive processes.

Let us now consider in some detail two early neuroeconomic experiments (Sanfey et al. 2003, de Quervain et al. 2004) that prepared the ground for many subsequent studies, and a more recent experiment that both measures the activation in areas of the brain and also introduces a neural-level intervention (Gospic et al. 2011).[7] We will argue that whereas the former imaging studies raise the same concerns about experimental artefacts as the purely behavioural experiments, implementing new experimental interventions can be an effective strategy in securing the right kind of independence.

Alan Sanfey et al. (2003) formulated an experiment in which the subjects play a modified UG while being scanned using fMRI. The normally anonymous game is modified so that the subjects play in random order both with a human partner, to whom they have been introduced, and with a computer. The players know whether they are playing with a human or a computer, although in reality the choice of the human proposer is pre-determined so that all the subjects receive the full range of offers in the responder role. The behavioural results are similar to those usually observed in UG experiments: low 'unfair' offers are rejected at a high rate. Interestingly, although the subjects seem to punish both humans and computers for unfair behaviour, the rate of rejection is significantly higher when they think they are playing with another human. Moreover, the fMRI indicates greater activation with unfair offers than with fair ones in the bilateral anterior insula, the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC). Furthermore, activation in these areas is significantly stronger when unfair offers are made by a human player as opposed to a computer, suggesting that the activation of these sites is not exclusively sensitive to monetary payoffs. Previous studies have shown that the bilateral anterior insula is associated with negative emotional states such as pain, hunger, disgust, anger and distress. In the above-mentioned experiments (Sanfey et al. 2003) the subjects with stronger anterior insula activation rejected a higher proportion of unfair offers, implying that anger, and the consequent wish to punish, is a direct

[7] See Fehr and Krajbich (2014) for an up-to-date review of the neuroeconomic literature on social preferences.

causal driver of the rejection of unfair offers. This is taken as evidence supporting the existence of social preferences (H_1). Sanfey et al. claim that the "results provide direct empirical support for economic models that acknowledge the influence of emotional factors on decision-making behaviour" (2003: 1758).

De Quervain et al. (2004) devised an experiment to test the hypothesis that people derive satisfaction from punishing norm violations. The subjects play a trust game (TG) while undergoing a PET scan. The TG is anonymous but the players know they are playing against other human players. The hypothesis that people are intrinsically motivated to punish norm violations implies that punishment behaviour, also known as altruistic punishment, should be associated with the activation of reward-related areas. To rule out alternative explanations of observed activation, De Quervain et al. (2004) introduced variation in terms of whether the abuse of trust in the game was intentional on the part of the other player or a product of random computer override, and whether punishment was purely symbolic, costly only to the abuser, or truly costly to both the abuser and the punisher. They integrated these variations into four treatment conditions: (i) Intentional and Costly, (ii) Intentional and Free, (iii) Intentional and Symbolic, and (iv) Non-intentional and Costly. The subjects' brains were scanned and differential activations were observed in five contrasts. As predicted, the activation of reward-related areas was greater when people had the opportunity (e.g. the difference between i) and ii)) or the desire (e.g. the difference between iii) and iv)) to punish. Moreover, those who felt greater satisfaction from punishing also invested more in it. These results, complemented with questionnaire results, are interpreted as confirming the claim that individuals do obtain satisfaction from altruistic punishment. The authors also conclude that the results "support recently developed social preference models, which assume that people have a preference for punishing norm violations…." (1258). In other words, they conclude that the results support H_1.

Any conclusion that these two studies provide support for the social-preference hypothesis should be made with caution, however. First, both studies rely on neuroimaging techniques without introducing neural-level interventions, and are therefore able, at most, to establish a correlation between activation and behaviour, and do not prove causation.[8] Second, in order for neuroeconomic experiments to play a role in the triangulation of social preferences, they should not share the same errors and biases with respect to the *motivational states* they are

[8] Ruff and Scott (2014) distinguish between *measurement techniques* (e.g. EEG, MEG, PET and fMRI) and *manipulation techniques* (brain stimulation and lesion studies). Measurement techniques cannot demonstrate that a particular region of the brain is necessary to a given cognitive function.

designed to track. If a behavioural result is, in fact, attributable to a feature of the experimental set-up that is not related to social preferences (such as if the subjects are simply wrong about the game), then the neuroeconomic data should be produced in a process that is not likely to be susceptible to the same kind of error. From this it follows that the probability of getting the kind of imaging data implied by the social-preference hypothesis should be independent of the results of the behavioural experiments, conditional on the hypothesis being true or being false. Let us consider the latter case: if social preferences were not the cause of altruistic behaviour such as punishment, would it be highly unlikely to obtain the kind of imaging data now produced in the neuroeconomic experiments?

As stated above, the principal reason to be concerned about the evidence for social preferences produced by behavioural-punishment experiments is that the punishment behaviour may be an artefact of the laboratory. Guala (2012) argues along similar lines, suggesting that the 'social' situation in such experiments is so sparse and the possible actions so constrained that it may well be that the subjects channel all their misgivings through the only available avenue, namely punishing the other player. If this were indeed the case, we would also expect to see activation of brain areas related to stress and anxiety (anterior insula) and 'competition' between these emotional and top-down systems (DLPFC), in other words the same areas that were found to show higher activity. Therefore, the probability of observing the kind of data actually observed in the neuroeconomic experiment, given the falsity of the social-preference hypothesis, is not sufficiently low. In this particular example, the experiments concerning neuroeconomic punishment might merely replicate the same artefactual punishment behaviour, and the imaging data might be insufficient to rule this out.

This highlights a more general problem: given that relevant observed neural activation interpreted as evidence tends to be indicative of general motivational mechanisms and affective states, some such activation would normally be expected to occur even if the behaviour in the laboratory were not attributable to social preferences. It is hard to construct imaging studies that are suitably independent as a means of detecting specifically social preferences if such studies remain purely observational on the neural level.[9] One way to achieve error-independence is to introduce causal controls on the level of neural mechanisms, thereby providing independent variation to some causal variable between social preferences and neural activation, which should not be affected if social preferences were artefacts of the

[9] Ross (2008) uses the label 'behavioural economics in the scanner' to refer to neuroeconomic experiments that merely replicate experiments in behavioural economics with the addition of information about the subjects' brain activity obtained from neuroimaging.

experimental apparatus (as per hypotheses H_2 and H_3). This could be achieved pharmacologically, using various ways of direct electro-magnetic stimulation and suppression, or in behavioural experiments on subjects with brain lesions.

To illustrate the difference between an intervention on the behavioural level and an intervention on the mechanistic neuro-level, let us look at a recent neuroeconomic study that includes a pharmacological intervention. Katarina Gospic et al. (2011) questioned the fact that Sanfey and colleagues' UG experiment only implicated cortical structures and not the amygdala, which has been robustly implicated in emotional responses elsewhere. They theorise that the evolutionary younger insula DLPFC-ACC network has to do with future-oriented decisions that take time, and that the relatively long decision-making window in the UG experiments masks the immediate gut reactions involving the evolutionarily older subcortical systems (especially the amygdala). In order to test this hypothesis, Gospic et al. set up a UG experiment with a shorter decision-making time. Crucially, they also added another treatment, benzodiazepine, which suppresses amygdala activity. They found that the fMRI did indeed show increased activation in the amygdala associated with the rejection of unfair offers in the UG with more rapid responses. The benzodiazepine-treatment group rejected unfair offers at a lower rate than the (placebo) control group, and the fMRI showed lower activation in the treatment group than in the control group. The researchers further theorise that their findings point towards the existence of an evolutionarily older "inequality aversion", perhaps developed to ensure an acceptable splitting of prey within a group, which does not rely on higher cortical processes.[10]

Although, strictly speaking, they propose a slightly alternative understanding of the decision mechanism underlying altruistic punishment (the hypothesised immediate gut reaction against inequality), Gospic et al. show how the introduction of new causal controls, in this case on the level of neural mechanisms, might result in the right kind of independence: if the hypothesis positing the involvement of the amygdala in the (rapid) rejection of unfair offers were false, and the behavioural phenomenon attributable to some other mechanism, then the observation of behavioural differences under the introduced treatment directly affecting the amygdala would have been highly unlikely. It is worth noting that the evidential value of the fMRI imaging in the study is secondary at most.

[10] They also point towards developmental evidence that small children exhibit similar tendencies without having the capacity for long-term planning or a developed theory of mind.

Above we claim that neuroscientific experiments can, in principle, be used to triangulate on phenomena relevant to economic models. Our story is congruent with that of Clarke (2014) in that the relevant phenomena are more likely to be psychological than purely behavioural, and that independent variation in these variables is important. The possibility of introducing neural-level causal interventions to increase the kind of independence required in triangulation is a powerful argument in favour of these experiments, but this epistemic value should be strictly distinguished from a common intuition that neuroscientific experiments have evidential value *purely by virtue of being neuroscientific*. The intuition is that by being closer to the physical decision-making mechanisms, neuroscientific evidence is more real than evidence from purely behavioural experiments or anthropological observation. Nor is it the case that neuroeconomics will provide triangulating evidence purely by virtue of offering *different* kinds of data if the respective methods are not suitably independent. Our worry is that both of these intuitions may bias the evaluation of the experimental evidence provided by neuroeconomics.

If these experiments really track stable, psychological decision-making mechanisms, then such mechanisms should also be stable across different behavioural designs. In other words, we would expect to see not only stable pro-social behaviour in different games in the aggregate but also stable pro-social tendencies (or associations of such behaviour) across different games within individual subjects (see e.g. Johnson et al. 2009). Varying the incentive structure, the way the game is presented to the players, anonymity and the information set should all be equally good (and significantly cheaper) ways of increasing the independence of experiments and distinguishing experimental artefacts from stable phenomena. It is also worth noting that the evidentially important new variable in Gospic et al.'s experiment was the pharmacological intervention, and the scanning merely served a secondary evidential purpose in shoring up the claim that the intervention did what it was supposed to do. The idea that imaging studies offer an independent alternative in terms of seeing motivating states is appealing, but as our account shows, the fact that a new way of detecting a phenomenon is *different* does not mean that it is *independent* in the way required for successful triangulation. If social preferences cannot be captured as stable individual psychological mechanisms across different games, using neuroscientific experiments to investigate what happens in the brains of the subjects in each game is of little help in establishing the external validity of the observed behavioural phenomena. We now discuss the plausibility of such extrapolation.

6. Social preferences in the wild

Observations of pro-social behaviour in the field are also taken to testify to the reality of social preferences (Camerer 2013). However, casual observation cannot reliably discriminate between the presence of social preferences and alternative explanations of pro-social behaviour. Alongside the countless scientific and lay observations is an emerging body of experimental work on pro-social behaviour conducted in a variety of cultural and social contexts around the world. Although the majority of these studies involve subjecting people to maximally controlled game-theoretical experiments –bringing the laboratory to different cultural and social contexts – attempts are made in some experiments to introduce an experimental element into the more ordinary flow of life.

When the laboratory is simply carried onto the field the resulting process of data generation is not sufficiently different to produce independent evidence. However, the appeal of many field experiments is precisely that the absence of laboratory conditions reduces the probability that the result is an artefact of some aspect of the laboratory situation. Natural field experiments – meaning that the subjects are not aware of being in an experiment and the environment is one in which they normally operate (Harrison and List 2004: 1014) – have been conducted to control for the possibility that pro-social behaviour is somehow attributable to the presence of the investigator and/or the artificiality of the laboratory situation (e.g. Winking and Mizer 2013, Balafoutas et al. 2014). Jeffrey Winking and Nicholas Mizer (2013), for example, carried out a natural field experiment in Las Vegas that was simple in design: a confederate was standing at a bus stop in front of a casino ostensibly about to get on a bus, while appearing to notice that he still had casino chips in his pocket. The confederate then offered these chips to an unsuspecting participant and suggested that the participant could share the chips with another confederate, thus realising an anonymous one-shot DG. As a comparison, a more standard DG field experiment with full knowledge of the experimental nature of the situation was conducted in a similar setting. The results were striking. Whereas the results of the comparison field experiment were broadly in line with the laboratory results, in the natural field experiment, which eliminated awareness of participating in an experiment, nobody donated anything. This seems to suggest that the 'please-the-investigator' hypothesis cannot be ruled out as a rival explanation of pro-social behaviour, at least in DG-like experiments – although other interpretations are possible.

7. From phenomena to theories

Now, for the sake of argument, let us suppose that the various behavioural, neuroeconomic and field experiments provide legitimate grounds on which to eliminate various error hypotheses about the causes of pro-social behaviour. Does the support that these diverse methods lend to the experimental phenomenon of social preferences carry over to general theories of human cooperation? Not necessarily. By way of illustration, let us focus on the relationship between *altruistic punishment*, an instance of social preference, and *strong-reciprocity theory*, a general theory of cooperation. Strong reciprocity theory holds that real-world cooperation is maintained by the willingness of individuals 1) to cooperate with cooperators even when behaving non-cooperatively might be more beneficial, and 2) to punish non-cooperators even when doing so is costly to themselves (altruistic punishment). The motivation to punish deviant or purely selfish behaviour (an instance of social preference) is taken to be the result of some form of group selection. Therefore, anthropological findings about small hunter-gatherer communities, which most closely correspond to the assumed Pleistocene (social) selection environment of our ancestors, are regarded as key evidence in favour of strong-reciprocity theory. The refusal of the responder to accept low offers in UGs is a laboratory manifestation of altruistic punishment, which is hypothesised to be a crucial mechanism for sustaining social coordination in everyday social life (e.g. Gintis, 2000; Gintis et al. 2005; Henrich et al. 2004).

Guala (2012) distinguishes between a 'narrow' and a 'wide' interpretation of the experimental evidence about altruistic punishment. Under the former the experimental evidence warrants claims about the existence of a laboratory phenomenon, namely that a certain kind of social preference causes observed altruistic punishment in laboratory experiments. Under the 'broad' interpretation, in turn, experimental evidence would warrant claims about the operation of the same mechanism in supporting cooperation in the wild. Guala argues that existing evidence warrants the narrow, but not the broad interpretation. The results of anthropological studies about small societies and studies on social dilemmas conducted by social and economic historians rather seem to suggest that cooperation in the wild is typically sustained by institutional mechanisms. Guala hypothesises that this is so because in the wild the costs of punishment for the punisher are so high that sustaining widespread cooperation with the decentralised mechanism of altruistic punishment would not be feasible. Note that Guala's critique does not concern the artificiality of the laboratory result as such. Altruistic punishment is a real laboratory phenomenon, according to Guala. The worry is rather that it is

not altruistic punishment but other institutional mechanisms that, in reality, explain human cooperation.[11]

Guala's points highlight the importance of separating phenomena-to-theory and data-to-phenomena inferences. The observation that altruistic punishment does not maintain cooperation in the wild speaks against strong-reciprocity theory, but does not necessarily cast doubt on the existence of social preferences as a real phenomenon. There is no tension between the varying determination of social preferences, which provides some support to strong-reciprocity theory, and the field evidence that speaks against it. In addition, strong-reciprocity theory is not the only theory of cooperation that implies the existence of social preferences (e.g. Bicchieri's theory of social norms, 2005).[12] Shoring up support for a phenomenon that all the competing theories explain obviously cannot help in discriminating between them. Triangulation is a matter of controlling for errors arising from particular methods of determination in relation to a specific hypothesis: the wider the scope of the hypothesis the more and varied are the errors that should be controlled for and the more diverse the evidence that should be brought to bear. It is partly because triangulation is frequently regarded as a strategy for the confirmation of general theories that its epistemic power has been questioned.

7. Conclusions

Bringing together evidence from diverse disciplines can increase the reliability of claims about the phenomenon under investigation. This is a matter of triangulation: the use of multiple and independent sources of evidence to determine, test or measure the same (aspect of a) phenomenon. However, the mere fact that evidence is brought in from different disciplines does not guarantee its independence, nor does methodological diversity. The epistemic value of triangulation is a red herring without an account of independence. We have argued that its epistemic value lies in cancelling out errors in *causal* inference to shore up claims about a phenomenon. Our account explains why and when converging evidence from different fields legitimately increases confidence in a phenomenon. In addition, although general scepticism about the incomparability of evidence and ontologies is overstated, we

[11] The difference between establishing the reality of the laboratory phenomenon and inferring that the phenomenon is exportable outside the laboratory roughly corresponds to the distinction between internal and external validity.
[12] To be precise, Bicchieri's theory of social norms also implies a different interpretation of social preferences, in other words of how individuals take others' utilities into account (cf. e.g. Bicchieri and Zhang 2012).

have shown that many of the concerns expressed about the feasibility of successful triangulation are legitimate. Whether or not they apply has to be ascertained in each case. Moreover, even successful triangulation is not a silver bullet that can solve the problem of extrapolation from the laboratory to the wild.

When applied to neuroeconomics in general and to the case of social preferences in particular, our account identifies the conditions under which convergence between evidence from neuroeconomic experiments and other sources increases confidence in the existence of a phenomenon. On the one hand, if a neuroeconomic experiment is suitably independent because experimental controls are introduced on the neural level, for example, then neuroeconomics contributes to the reliability of inferences concerning the causes of pro-social behaviour in the lab. On the other hand, the right kind of independence might well be obtained by means of varying the experimental paradigm on the behavioural level. The requirement of independence applies to the evidential role of neuroscience with respect to our theories of social behaviour in general. Assessing the role of neuroscientific evidence in the social sciences requires consideration of the specific causal processes through which the data are generated, as well as of the specific claim for which they constitute evidence. We therefore conclude that both wide-ranging scepticism and unconditional optimism with regard to the potential of neuroscience to triangulate on social scientific phenomena are unwarranted.

References

Bicchieri, C. (2005) *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., & Zhang J. (2012). An Embarrassment of Riches: Modeling Social Preferences in Ultimatum Games. in U. Maki (Ed.) : Elsevier.

Blaikie, N. (1991) "A critique of the use of triangulation in social research", *Quality and Quantity* 25: 115-136.

Bogen, J. and Woodward, J. (1988) "Saving the phenomena." *The Philosophical Review* 47(3) 303-352

Burton-Chellew, M. and West, S. (2013) Prosocial preferences do not explain human cooperation in public goods games, *PNAS* 110: 216-221.

Camerer, C. (2000) Prospect Theory in the Wild: Evidence from the Field.chap. 16 in D. Kahneman and A. Tversky, eds., *Choices, Values, and Frames*, Cambridge University Press.

Camerer, C. (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton Univ. Press.

Camerer, C. (2013) Experimental, cultural, and neural evidence of deliberate prosociality, *Trends Cogn Sci*. 17: 106-107.

Clarke, C. (2014) "Neuroeconomics and confirmation theory". *Philosophy of Science* 81(2): 195-215

Craver, C. and Alexandrova, A. (2008) No revolution necessary: neural mechanisms for economics. *Economics and Philosophy* 24 (3): 381-406

De Quervain, D.J.-F., Fischbacher, U., Treyer, V. Schellhammer, M., Schnyder, U., Buck, A., Fehr, E. (2004) "The neural basis of altruistic punishment." *Science* 305: 1254-1258

Fehr, E. and Krajbich, I. (2014) "Social Preferences and the Brain", in Glimcher, P. and Fehr, E. (eds.) *Neuroeconomics: Decision Making and the Brain* (2nd ed.), Academic Press. 193-218.

Fitelson, B. (2001) "A Bayesian account of independent evidence with applications." Philosophy of Science 68: S123-S140

Fumagalli, R. (2013) "The futile search for true utility." *Economics and Philosophy* 29 (3): 325-347

Gigerenzer, G. and Gigerenzer, T. (2005) "Is the Ultimatum Game a three-body affair?" *Behavioral and Brain Sciences* 28(6): 823-824

Gintis, H. (2000) "Strong reciprocity and human sociality" Journal of Theoretical Biology 206: 169-179

Gintis, H. Bowles, S. Boyd, R. and Fehr, E. (eds) (2005) *Moral Sentiments and Material Interests*. MIT Press

Glimcher, P. (2011) *Foundations of Neuroeconomic Analysis*. Oxford University Press

Guala, F. (2012) "Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate?" *Behavioral and Brain Sciences* 35: 1-59

Gul, F. and Pesendorfer, W. (2008) "The case for mindless economics." In Caplin and Schotter 2008, 3-39

Hammersley, M. (2008) “Troubles with triangulation.” In: Bergman, M.M. (ed) *Advances in Mixed Method Research*. London, Sage: 22-36

Hausman, D. (2012) *Preference, Value, Choice, and Welfare*. New York: Cambridge University Press.

Henrich, J., Boyd, R. Bowles, S., Camerer, C. Fehr, E. Gintis, H. (2004) *Foundations of Human Sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press

Hudson, R. (2013) *Seeing Things. The Philosophy of Reliable Observation*. Oxford University Press

Johnson, T., Dawes, C., Fowler, J., McElreath, R. and Smirnov, O. (2009) The role of egalitarian motives in altruistic punishment, *Economics Letters* 102: 192-194.

Kuorikoski, J. and Ylikoski, P. (2010) “Explanatory relevance across disciplinary boundaries. The case of neuroeconomics.” *Journal of Economic Methodology* 17: 219-228

Poldrack, R. (2006) “Can cognitive processes be inferred from neuroimaging data?” *Trends Cogn Sci*. 10(2): 59-63.

Rick, S. (2011) Losses, gains, and brains: Neuroeconomics can help to answer open questions about loss aversion. *Journal of Consumer Psychology* 21: 453–463.

Ruff, C.C. and Huettel, S.A. (2014) “Experimental methods in cognitive neuroscience.” in Glimcher, P. and Fehr, E. (eds.) *Neuroeconomics: Decision Making and the Brain* (2nd ed.), Academic Press, 77-108.

Ross, D. (2008) “Two styles of neuroeconomics.” *Economics and Philosophy* 24: 473-483.

Sanfey, A.G., Riling, J., Aronson J.A, Nystrom, L.E, Cohen J.D. (2003) “The neural basis of economic decision-making in the ultimatum game.” *Science* 300: 1755-1758

Schupbach, J. (2015) “Robustness, Diversity of Evidence, and Probabilistic Independence.” In Mäki, Ruphy, Schurz and Votsis (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*. Springer: 305-316

Stegenga, J. (2009) “Robustness, discordance, and relevance.” *Philosophy of Science* 76: 650-661

Stegenga, J (2012) “*Rerum concordia discors*: robustness and discordant multimodal evidence.” In L. Soler et al. (eds.) *Characterizing the Robustness of Science*, Boston Studies in the Philosophy of Science 292

Wimsatt, W. (1981) “Robustness, reliability, and overdetermination.” In M. Brewer and B. Collins (eds.) *Scientific Inquiry in the Social Sciences*, San Francisco: Jossey-Bass: 123-162

Woodward, J. (2009) “Experimental investigations of social preferences”. In Kincaid and Ross (eds.) *The Oxford Handbook of Philosophy of Economics*.

Zizzo, D.J. (2010) “Experimental demand effects in economic experiments.” *Experimental Economics* 13: 75-98