Published in *Journal of Economic Methodology* 17: 219-228.

EXPLANATORY RELEVANCE ACROSS DISCIPLINARY BOUNDARIES – THE CASE OF NEUROECONOMICS

Jaakko Kuorikoski[a*] and Petri Ylikoski[b]

[a]*University of Helsinki, Finland;* [b]*University of Tampere, Finland*

Abstract:

Many of the arguments for neuroeconomics rely on mistaken assumptions about criteria of explanatory relevance across disciplinary boundaries and fail to distinguish between evidential and explanatory relevance. Building on recent philosophical work on mechanistic research programmes and the contrastive counterfactual theory of explanation, we argue that explaining an explanatory presupposition or providing a lower- level explanation does not necessarily constitute explanatory improvement. Neuroscientific findings have explanatory relevance only when they inform a causal and explanatory account of the psychology of human decision-making.

Keywords: neuroeconomics; explanatory relevance; levels of explanation; interdisciplinarity; mechanisms

1. Introduction

The promise of neuroeconomics is the possibility of achieving better integration of knowledge across scientific fields. Many arguments for and against neuroeconomics revolve around ideas about explanation, unification, and division of labour between disciplines. Many

* Email: jaakko.kuorikoski@helsinki.fi

of these arguments are confused because they rely on mistaken assumptions about criteria of explanatory relevance across disciplinary boundaries, and they also fail to distinguish between evidential and explanatory relevance. Based on the mechanistic account of neuroscientific explanation (Craver 2007; Bechtel 2008) and the contrastive counterfactual theory of explanation (Woodward 2003; Ylikoski 2007), we argue that the mistaken attributions of explanatory relevance are mainly due to two false ideas about explanation: that explaining an explanatory presupposition automatically improves the original explanation and that a lower-level mechanistic explanation is always superior to a higher-level explanation. It is useful to make a distinction between two quite different programmes within neuroeconomics: the first attempts to use neuroscientific findings to explain economic phenomena (behavioural economics in the scanner, as Ross 2008 calls it). The second uses the mathematics developed for economic equilibrium analysis to develop models for computational neuroscience (neurocellular economics). Our discussion will be limited to the former.

Our main argument is that the idea of a direct connection between economics and neurosciences is misguided. These two fields can only be integrated via psychological theories of decision-making: neuroscientific findings have neither explanatory nor evidential relevance to economics unless these findings are interpreted in the light of substantial psychological theories. In our view, such integration is badly needed: economics needs a causally explanatory psychology of decision-making to support its behavioural assumptions and the future advancement of neuroscientific understanding of decision-making is dependent on the development of psychological theories. This idea goes against (some of) the rhetoric of neuroeconomics, which attempts to give the impression that neuroscience is directly relevant to economics (e.g., attempts to show that utility is a physiological entity). A closer look at

neuroeconomists' 'recent neuroscientific theories' (cf. Camerer et al. 2005) shows that these are psychological and social psychological theories, most of which were developed several decades ago. We argue that recognition of the proper role of psychological theory is crucial for the future intellectual respectability of neuroeconomic research. The current practice of interpreting neuroscientific findings in light of informal, common-sense psychology is not acceptable, as the localization of components of decision-making processes makes sense only in the context of substantial psychological theory. Without such a theory the localization hypotheses do not have any evidential or explanatory value.

Although we hold that neuroscientific evidence is not directly explanatorily relevant for economics, we do not share Gul & Pesendorfer's (2008) view that economics should be kept completely isolated from other disciplines. Sensible division of cognitive labour does not imply that economics should be a separate science. In our view, economics would benefit from closer integration with other sciences. The assumptions and explanations of economics should be consistent with the findings and theories of the various sciences that study human behaviour across different levels of organization. The crucial question is into which sciences should economics be proximately integrated? We will address this question by considering the conditions in which knowledge of brain mechanisms underlying economic behaviour could lead to an explanatory improvement *with respect to the economic phenomena*. Our thesis is that neuroscientific findings provide evidence primarily for psychological research, and neuroscientific findings are relevant for economics only when mediated via substantial psychological theories.

In this paper, we will not address the methodological challenges in the use of neuroscientific data to support localization hypotheses. There are general inferential problems with these hypotheses (Uttal 2001; Glymour 2001) as well as more local technical challenges. For example, the localisation claims made on the basis of fMRI-imaging studies face significant

problems owing to the temporal and spatial resolutions of the data and the inability of the BOLD signal to distinguish between active processing and neuromodulation (Logothetis 2008). There are also problems with experimental design, sample sizes and statistics used (Harrison 2008). Our concern in this paper will be the explanatory relevance of the neuroscientific findings that pass these methodological hurdles.

2. Explanatory relevance across levels of mechanisms

Many advocates of neuroeconomics claim that they are in the business of opening the black box of the decision-making agent and are thus advancing a mechanistic explanation of individual economic behaviour (e.g., Camerer et al. 2005). Opening up black boxes and finding explanations in terms of "lower-level" mechanisms is often regarded as a self-evident improvement in our explanatory knowledge. Our aim in this section is to replace this intuitive view with an explicit theory of explanation and to see under what conditions and to what extent opening black boxes can be regarded as an explanatory improvement *within economics*.

According to the contrastive counterfactual theory of explanation (Woodward 2003; Ylikoski and Kuorikoski 2009), explaining a phenomenon amounts to exhibiting the factors it depends on.[1] Dependence can be analysed using counterfactual conditionals: *A* depends on *B* if it is true that if *B* had been different, then *A* would have been different as well. Furthermore, most ordinary explanations expressed in natural language are ambiguous and can be made more precise by explicating *contrasts* for the thing to be explained (the *explanandum*) and the explanatory factor (the *explanans*). This is based on the fact that explanations do not relate events or phenomena as a whole. An individual explanation always addresses only specific aspects of phenomena, which are conceptualised in some specific manner. Explicating the contrast is a way of making this conceptualisation explicit (Ylikoski 2007). Explanations can

be considered as answers to questions in the form Why is it the case that A rather than A*. The answers to these questions also have a contrastive form: B rather than B* being the case is the explanation for the occurrence of A rather than A*. More generally, explanations can be taken to relate variables. The formulation of explanations in terms of variables explicates the implicit space of alternative possibilities in which the explanations are considered (the contrast classes), thus making both the *explanans* and the *explanandum* more precise.

What makes a piece of information explanatory rather than purely descriptive is that the exhibited dependency enables one to infer to counterfactual situations beyond what actually happens. This ability to answer *what-if-things-had-been-different* -questions (*what if* -questions for short) can be taken as the core of the intuitive notion of understanding (Ylikoski 2009). Our theory of explanation thus rests on the basic idea that the more inferences you can make with the help of the explanatory information (including inferences about things beyond what actually happened), the better off you are in terms of explanatory understanding. A crucial insight provided by the contrastive counterfactual theory is that causal explanatory knowledge is necessary for manipulation. Merely descriptive knowledge (knowledge of regularities) is sufficient for non-manipulative, passive prediction, but knowledge about causal dependencies underlying any manifest regularity is necessary when predicting the behaviour of the system when it is subjected to exogenous changes, such as goal-directed manipulation. (Woodward 2003; Ylikoski and Kuorikoski 2009.)

We are now in a position to argue that an explanation is not necessarily improved when the *explanans* is itself explained. To see why this is the case, one has to see the difference between explaining a phenomenon and explaining an *explanatory presupposition* of that explanation. Once the contrast class for the *explanandum* is set, there is (usually) a determinate fact of the matter of what its explanation is. When we know what the *explanandum* variable is dependent on and what form this dependency takes, there is nothing

more to be explained about this particular contrastive *explanandum*. For example, once we know that whether a player rejects rather than accepts a specific offer in the ultimatum game depends on the intensity of the player's social preferences (if there is such a thing), the explanation-seeking 'why' question is answered. Further explaining the *explanans*, whether it be causally explaining the value of the *explanans* variable (why the subject came to have such social preferences) or constitutively explaining the form of the explanatory dependency (how the decision making mechanism is realised), is not an explanatory improvement with respect to the original *explanandum*. Even if we knew that the social preferences and the associated deliberation mechanism were (partly) realised by the neurons in the ventromedial PFC Broadman areas 10 and 11, this information would not add to the original explanation of why rejection rather than acceptance, since the activation of the neurons mentioned would make the same difference as the social preferences. Of course, discovering new explanations increases our overall understanding of the phenomenon of decision making, but it may do little to advance our understanding about the original domain of phenomena of interest (e.g., economic phenomena involving altruistic punishment). Explaining such explanatory presuppositions may have evidential virtue in that our knowledge of the original *explanans* is more secure, owing to our new findings (discovering the neural mechanisms realising social preferences is evidence for their existence), but evidential and explanatory virtues are not the same thing (Ylikoski and Kuorikoski 2009).

Nor is it the case that a 'lower-level' explanation that replaces the original explanation is always to be preferred to a higher-level one. The details of the lower-level mechanisms that realise the upper-level variables are explanatory irrelevant for the higher-level *explanandum,* if changes in the lower-level detail do not make a difference to the value of the original *explanandum* variable. Even if we had secure knowledge about the lower-level mechanisms, and in the case of neuroeconomics we certainly do not, this knowledge would be irrelevant if

it did not enable further *what if* -inferences concerning the original economic *explananda*. For example, Fehr and Camerer (2007) argue that multiple imaging studies showing activation of dorsal and ventral striata provide evidence for a hedonistic interpretation of social preferences, i.e., that a reward sensation has a causal role in the production of pro-social behaviour. Outside the highly unusual situations in which this particular causal link in the etiology of pro-social behaviour is broken, the knowledge of whether something happens in the dorsal or the ventral striatum does little to improve our economic explanations. This is because the mechanistic neural detail does not make a further difference with respect to the social and economic phenomena to be explained, given that we already know the causal role of social preferences. The neural information simply does not allow us to make any new relevant *what if* -inferences about economic phenomena.

The lesson to be learned from this is that *explanations find their own level*. There is no privileged level of explanation from which every other level of description or organisation inherits whatever explanatory power they have. Higher-level explanations do not inherit their explanatory qualities from lower-level descriptions; they are explanatory because of the counterfactual information provided by the explanatory dependency, not as mere placeholders for a future 'true' lower-level explanation (Ylikoski and Kuorikoski 2008).

3. The case for a missing level

We are now in the position to argue that the individual-level variables that are explanatorily relevant to economic phenomena are (at least usually) psychological, and that there is likely to be no viable shortcut from the neural level to economic phenomena. By the term psychological, we refer to causal processes mediating agents' inputs and outputs that are described and individuated using informational or representational vocabulary. The neuroscientific findings make sense only in the light of substantial psychological theory. It is

not that brain processes are irrelevant to the production of behaviour, but that knowledge of such causal relationships does not serve the theoretical or practical purposes of economics. The fact that neuroscientific data can serve as evidence for psychological states and processes is not a basis for the direct integration between neurosciences and economics. Our argument is that knowledge about direct dependency relationships between neuroscientific variables and economic phenomena would be of little theoretical or practical use. Even if we could establish the existence of stable dependency relationships between neural phenomena and economic phenomena in a laboratory setting, these dependencies would almost certainly 1) have very limited applicability beyond the laboratory, 2) have only limited counterfactual range ("explanatory power") and 3) be very difficult to integrate into other bodies of explanatory knowledge.

1) Suppose that in some experimental setting there is a stable, invariant relation of dependence between increased activity in the nucleus accumbens or in the anterior insular cortex and the subject exhibiting risk-seeking or risk-averse behaviour (Kuhnen and Knudson 2005). Although in a very limited way, the difference in brain activity can be said to explain the difference in attitude to risk in the context of the specific experimental decision problem, the explanation hardly goes beyond the trivial point that the behaviour is (to a large extent) controlled by processes in the brain. This explanation does not improve our understanding of *economic* phenomena, since it is hard to think of a scenario in which we would have access to data on people's brain area activation, but not to data about the behaviour of the economic agents. The observable behaviour and the psychological variables that can be inferred from the behaviour screen off the neural details, so to speak. If we explain a market outcome on the basis of activations of anterior insular cortices, we are only pointing out (at most) that the agents were, in fact, risk averse. Of course, there is a sense in which the mass activation of

anterior insular cortices of investors does explain the fall in stock prices in the autumn of 2008, but this explanation strikes us as a bit silly and pretentious, and rightly so.

Naturally, neuroimaging data could in principle be used as evidence for values of psychological parameters that could be relevant outside the laboratory and could thus be used to improve explanations of economic phenomena. The trouble is that most laboratory experiments that correlate differences in brain activation *across* subjects to psychological or behavioural differences are highly indexical to the experimental set-up and therefore have poor external validity. For example, we cannot reliably use imaging data associated with risk-taking behaviour of a single subject in a specific experimental decision situation to predict her behaviour in other decision situations, since risk-taking behaviour has turned out to be highly context specific (Platt and Huettel 2008).

2) The dependency between localizable brain activity and risk-taking also has very limited explanatory power in the sense that we have little knowledge about how risk-taking would be affected if the explanatory variable (brain activity) were to take slightly different values. At least in the light of present knowledge, there is no systematic dependency between a brain-area activation intensity and behaviour beyond the binary case. The range of counterfactual *what if* -questions that can be answered on the basis of the dependence is consequently very limited: we cannot infer to things beyond the rough expectations of the mean level of risk-taking on the basis of rough averages of brain-area excitation. In contrast, if we model the choice of an agent in psychological (cognitive) terms, for example, as depending on what the agent values, knows and how the decision situation is framed, then we can reasonably answer a broader range of *what if* -questions concerning possible alterations in the agent's valuations, knowledge and how the relevant information is presented. If we model a choice as being dependent on whether the agent is risk-seeking or risk-averse and if we know the information set the agent has at her disposal, then we can expect to answer *what if* -questions concerning

possible levels of risk-aversion and possible changes in the agent's information set. Psychological variables give us more inferential power than neural variables and thus more understanding about the economic phenomena.

The above points also make neuroscientific variables poor causal variables in that we have few possibilities to intervene on them in order to influence economic behaviour (cf. Woodward 2003). If we were interested in any causal handles that might be used to influence the market outcome, say, to be more like the one following from widespread risk-averse behaviour, any identified neural variables would not be of much use in practise. Although there are some studies identifying neural variables that might be directly intervened on in practise, e.g., introducing oxytocin into breathing-air to change people's behaviour in trust-related games, these interventions would be ethically problematic, to say the least.[2] Since explanations find their own level, we should discard the idea that some level of organisation or description is explanatorily privileged just because "the true causal work" happens there (i.e., the level is in some common-sense physical) or is the stuff of real hard science (e. g., carried out by people wearing white jackets and using sophisticated technical apparatus). After these ideas are discarded, it becomes hard to see what practical use the direct neuro-economic dependencies could be.

3) An immediate response to the previous accusation of practical uselessness would be to point out that science is first and foremost in the business of providing a theoretically unified picture of the world and that the neuroscientific level is theoretically unifying (e.g., Fehr and Camerer 2007). As Park and Zak (2007, p. 389) put it, "[neuroscientifically] augmented economic models will also likely include results from sociology, anthropology, psychology, and other fields. These can usefully be incorporated into economic models through the common pathway of the brain." We believe the opposite: direct neuro-economic dependencies are hard to integrate with existing theoretical knowledge. Economic

phenomena are best thought of as depending on variables that are not constituted solely by the intrinsic properties of an individual agent, let alone of an individual brain. Instead, decision variables are relational, and cognition involved in economic activities is an instance of extended cognition.

Since neuroeconomics is in the business of opening the black box of an individual decision maker, it already presupposes that it is the properties of individuals that are most relevant for economic theory. But for economic theory, it is often a mistake to conceptualise these variables as intrinsic properties of individual decision makers. It is widely acknowledged that straightforward localisation of a cognitive function to a brain area simply on the basis of imaging data is fallacious. What is not yet as widely recognised is that the localisation of economically relevant decision variables inside individual heads may not be much more sensible. This is because of two oversights in economic theory. First, as social psychological research strongly suggests, many of the variables that determine economic outcomes are relational properties of the decision situation, not intrinsic properties of the weighed alternatives or of the decision maker (Ross and Nisbett 1991).[3] If we restrict our attention to what happens inside a brain, we might lose most of the significant variables affecting economic phenomena. Second, as sociological research has recently emphasised, a major part of the functioning of the markets is in the material means of conducting trade. Economic decision making employs external tools of cognition extensively. (e.g., MacKenzie 2009.) These two considerations make any dependencies between neural and economic variables highly context dependent, more of a fluke of the particular circumstances than something that could be systematically integrated into a larger body of theory.

4. Integrating knowledge the mechanistic way

In this section we argue that the proper way to conceptualise the integration of neuroscience and economics is as parts of a mechanistic research programme. We contrast the model of a mechanistic research programme on the one hand, to forms of explanatory unification appealed to by some advocates of neuroeconomics, and, on the other hand, to appeals for hermetic insulation of economics as a separate science of choice aggregation made by some prominent deniers of neuroeconomics (Gul and Pesendorfer 2008).

A typical mechanistic programme of opening black boxes proceeds according to the heuristics of *functional decomposition and localisation* (Bechtel and Richardson 1993; Craver 2007). First, the different phenomena that the system of interest exhibits are differentiated. Then the phenomenon of interest is *functionally decomposed* in the sense of being analysed into a set of possible component operations that would be sufficient to produce the phenomenon. One can think of this step as thinking of a preliminary set of simple functions that, taken together, would constitute a more complex input-output relation (the system-level phenomenon). The system is also s*tructurally decomposed* or analysed into a set of component parts. The final step is to try to *localise* the component operations by mapping the operations unto appropriate structural component parts. The primary meaning of localisation here is the pairing of operations and parts, not (necessarily) that of locating something in physical space. The idea is thus to first think of what kinds of more basic properties or behaviours could, taken together, result in the *explanandum* behaviour and then try to determine whether the system is in fact made of such entities that can do the jobs required.

If and when the above cannot be done, then the fault may lie either in the manner of the functional decomposition or of the structural decomposition (or both), and these may then

have to be rethought. If a discovery about the lower-level realising mechanism means that the system cannot perform a certain function in a way that was assumed in the initial functional decomposition, then the functional decomposition may have to be altered. Conversely, if our functional decomposition seems otherwise sound (fits well in the pattern of causal explanations in higher-level terms), but is hard to reconcile with what we believe about the realising mechanism, then we may have to rethink the way we have conceptualised the components of the realising mechanism. In the end, even the identification of the target phenomenon or system may have to be revised. Mechanistic research does not eliminate higher levels of organisation nor does it treat higher or lower levels as sacrosanct. The goal is to find mutually consistent dependencies between things at different levels of organisation.

This mechanistic research programme provides the proper context for evaluating the localisation hypotheses presented by neuroeconomics. Functional localisation plays an important role in mechanistic research, but every neuroscientist acknowledges that a mere functional localisation does not by itself explain anything (Cacioppo et al. 2003; Henson 2005; Coltheart 2006). Observing that something happens in a specific area of the brain when a subject exhibits hyperbolic discounting or altruistic punishing behaviour does not by itself explain the behaviour. As was argued in the previous section, the variables that are explanatorily relevant for economic phenomena are psychological, and neuroscientific evidence is explanatorily relevant for psychological variables if and only if it constrains or informs the functional decomposition of the psychological theory. Neuroscientific discoveries can and should be relevant for economics, but only via the psychological level.

The trouble is that so far most neuroeconomic results do not have such implications or they simply demonstrate behavioural and cognitive deviations from the standard economic modelling assumptions that have been well known for decades (on the basis of behavioural studies). Is it really that surprising that people exhibit aversion to social betrayal beyond

monetary loss, that decisions dependent on and affecting the actions of other people involve affective as well as cognitive factors, that most brain processes are unconscious or that most of us buy more at the grocery store when we are hungry? Finding coherence between activations of specific brain areas and such psychological and behavioural phenomena does provide some additional evidence for the existence of such phenomena, but this evidential import is still distinct from explanatory import. However, there is a telling asymmetry even in this evidential use: neuroscientific findings are so far only used for confirmatory purposes. It is generally regarded as an achievement that imaging studies can distinguish between two cognitive processes at all. Finding such contrasts is regarded as supporting hypotheses that postulate such differences. However, the failure to find such contrasts is not regarded as disconfirming such hypotheses. Thus, even the triangulating function of neuroscientific findings is not yet taken to be all that reliable.

Integration of knowledge according to the mechanistic picture is tightly constrained by the actual causal and constitutive dependencies in the system investigated: a lower-level finding is explanatorily relevant if the new knowledge allows us to make new accurate *what if* -inferences about the behaviour of the system. The validity of these inferences depends on the causal facts in question: knowledge of a neuroscientific mechanism is explanatorily relevant to cognition or behaviour if we can use it to answer new contrastive *what if* -questions concerning cognition or behaviour. There is no *a priori* guarantee that the form of these new inferences resembles patterns of inference applicable to the original system level. Even if we can model choice behaviour as maximization of some function, it does not mean that whatever mechanism is (partially) causally responsible for the choices should or could be modelled as maximizing something – just as the fact that we can model the functional role of the cerebellum using adaptive filter models does not mean that we can do the same to the individual neurons of the cerebellum. In contrast, much of the unificatory appeal of

economics in general and neuroeconomics in particular is in the perceived universal applicability of the abstract principles of constrained optimization. If familiar patterns of reasoning are applicable to a set of new phenomena, then formulating explanations for these novel phenomena is *easier* (cognitively less demanding), but this alone is not constitutive of explanatory relevance (Ylikoski and Kuorikoski 2009).

Difficulties in conceptualising the proper integration of knowledge may also contribute to misplaced attributions of explanatory relevance in other ways. Since there is no explicit and openly discussed set of standards for explanatory relevance across disciplinary boundaries, it is easy for researchers to fall victim to various *illusions of depth of understanding* (Skolnick Weisberg et al. 2008; Ylikoski 2009) when interpreting results outside their specialty. This is true for economists trying to evaluate the relevance of neuroscience as well as for the larger public (and the rest of academia) trying to evaluate the relevance of neuroeconomics. An illusion of depth of understanding simply means that a person overestimates the detail, coherence and extent of her understanding. Based on a subjective sense of understanding, the scientist overestimates the number of correct *what if* -inferences about the phenomenon. These problems are especially prominent in cross-disciplinary explanatory endeavours, as scientists' sense of understanding is usually poorly calibrated for knowledge outside their own specialities. (Ylikoski 2009.)

The danger of illusory understanding is especially prominent in the case of neuroeconomics, since there is some experimental evidence that neuroscientific detail is, for some reason, especially conducive to misplaced attribution of explanatory relevance. Weisberg et al. (2008) have shown that inclusion of irrelevant neuroscientific details in a research report makes the report more convincing to non-experts. Even the simple inclusion of images of the brain in a paper or on the title page has similar effects on the evaluation of explanatory relevance (McCabe and Castel 2008). The trouble with neuroeconomics is precisely that

many in the intended audience are *not* experts with respect to neuroscience and hence are susceptible to this effect. This experimental evidence strongly indicates that intuitions based on the sense of understanding should be replaced by explicit criteria of explanatory relevance in the assessment of the relevance of neuroscientific findings to economics.

5. Conclusions

The discovery of neural mechanisms underlying economic decision making does not automatically improve economic explanations. Explaining an explanatory presupposition or providing a lower-level explanation does not constitute explanatory improvement unless the neuroscientific information provides grounds for making new *what if* -inferences concerning the original economic *explananda*. Neuroeconomics should be seen as a part of a mechanistic research programme, a mosaic of interconnected and mutually consistent set of explanations across different levels of organisation, not as a driver of grand unification of the social sciences. This implies that neuroeconomic data are explanatorily relevant only when they inform a causal and explanatory account of the psychology of human decision-making. So far they have not done so.

Acknowledgements

We thank the audiences of the Models, Mechanisms, and Interdisciplinarity -workshop in Helsinki and Neuroeconomics: Hype or Hope –conference in Rotterdam for their valuable comments.

References

Ariely, D. and Norton, M. I. (2008), How Actions Create – Not Just Reveal – Preferences, *Trends in Cognitive Sciences* 12 (1): 13-16.

Bechtel, W. & Richardson, R. C. (1993), *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, Princeton, NJ: Princeton University Press.

Bechtel, W. (2008), *Mental Mechanism. Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Cacioppo, J. T., Berntson, G., Lorig, T. S., Norris, C. J., Rickett, E. and Nussbaum, H. (2003), Just Because You're Imaging the Brain Doesn't Mean You Can Stop Using Your Head: A Primer and Set of First Principles, *Journal of Personality and Social Psychology* 85: 650-661.

Camerer, C., Loewenstein, G. and Prelec, D. (2005), Neuroeconomics: How Neuroscience Can Inform Economics, *Journal of Economic Literature* XLIII (March 2005): 9-64.

Coltheart, M. (2006), What Has Functional Neuroimaging Told Us about the Mind (So Far)?, *Cortex* 42: 323-331.

Craver, C. (2007), *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, New York and Oxford: Clarendon Press.

Fehr, E. and Camerer, C. (2007), Social neuroeconomics: the neural circuitry of social preferences, *Trends in Cognitive Sciences* 11: 419-427.

Glymour, C. (2001), *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, Mass.: The MIT Press

Gul, F. and Pesendorfer, W. (2008), The Case for Mindless Economics. In A. Caplin and A. Schotter (eds), *Foundations of Positive and Normative Economics*. New York: Oxford University Press.

Henson, R. (2005), What can functional neuroimaging tell the experimental psychologist?, *The Quarterly Journal of Experimental Psychology* 58A: 193-233.

Kuhnen, C. and Knutson, B. (2005), The Neural Basis of Financial Risk Taking, *Neuron* 47: 763-770.

MacKenzie, D. (2009), *Material Markets: How Economic Agents are Constructed,* Oxford and New York: Oxford University Press.

McCabe, D. P. and Castel, A. P. (2008), Seeing is believing: the effect of brain images on judgments of scientific reasoning, *Cognition* 107 (1): 343-352.

Park, J. W. and Zak, P. J. (2007), Neuroeconomics Studies, *Analyse & Kritik* 29: 47-59.

Platt, M. L. and Huettel, S. A. (2008), Risky business: the neuroeconomics of decision making under uncertainty, *Nature neuroscience* 11: 398-403.

Ross, L. and Nisbett, R. (1991), *The Person and the Situation: Perspectives of Social Psychology*, Philadelphia PA: Temple University Press.

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

Uttal, W. R. (2001), *The New Phrenology. The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, Mass.: The MIT Press.

Skolnick Weisberg, D., Keil, F. C., Goodstein, J., Rawson, E. and Gray, J. R. (2008), The Seductive Allure of Neuroscience Explanations, *Journal of Cognitive Neuroscience* 20: 470-477.

Ylikoski, P. (2007). The Idea of Contrastive *Explanandum*, in Persson, J. and Ylikoski, P.

(eds): *Rethinking Explanation. Boston Studies in the Philosophy of Science* 252, Springer: 27-42.

Ylikoski, P, (2009), The illusion of depth of understanding in science, in De Regt, Leonelli and Eigner, (eds): *Scientific Understanding: Philosophical Perspectives*. Pittsburgh University Press: 100-119.

Ylikoski, P. and Kuorikoski, J. (2008), Intentional Fundamentalism, in Hieke & Leitgeb (eds.): *Reduction and Elimination in Philosophy and the Sciences - Papers of the 31st International Wittgenstein Symposium Vol XVI*, Kirchberg am Wechsel, Austria: Austrian Ludwig Wittgenstein Society, 405-407.

Ylikoski, P. and Kuorikoski, J. (2009), Dissecting Explanatory Power, *Philosophical Studies*, forthcoming.

---

[1] Limitations of space mean that we cannot here provide additional arguments for our adopted theory of explanation. For these, see the cited references.

[2] Park and Zak (2007, p. 390) state that ”the role of oxytocin, and more generally empathy, in building trust has clear implications for institutional design to increase trade”. If the reference to oxytocin is removed, the statement does not go beyond Adam Smith. If the reference to empathy is removed, the statement suggests a novel and frightening approach to institutional design.

[3] It is again important to note that this context-dependency does not make neuroscience irrelevant for understanding human behaviour in general; quite the contrary (see Ariely and Norton 2008).