For a Few Neurons More.
Tractability and Neurally-Informed Economic Modelling

Abstract There continues to be significant confusion about the goals, scope and nature of modelling practice in neuroeconomics. This paper aims to dispel some such confusion by using one of the most recent critiques of neuroeconomic modelling as a foil. The paper argues for two claims. First, currently, for at least some economic model of choice behaviour, the benefits derivable from neurally-informing an economic model do not involve special tractability costs. Second, modelling in neuroeconomics is best understood within Marr's three-level of analysis framework and in light of a co-evolutionary research ideology. The first claim is established by elucidating the relationship between the tractability of a model, its descriptive accuracy, and its number of variables. The second claim relies on an explanation of what it can take to neurally-inform an economic model of choice behaviour.

1 Introduction

Should economic models of choice behaviour be informed by findings and concepts from the cognitive and biological sciences? This question has received considerable attention since the rise of neuroeconomics. So far, most critics of neuroeconomics have focused either on principled, a priori reasons for why results and concepts from cognitive neuroscience should not inform economic models of choice, or on evidential issues, which make it at least premature to bring results from cognitive neuroscience to bear on economic modelling.

Gul and Pesendorfer ([2008]), for example, argue that neurophysiological data are irrelevant in principle to economic models, as economic models are not designed to explain those data. Rubinstein ([2008]) observes that to date brain studies have not produced any

novel relevant insight for economics. Harrison ([2008]) and Harrison and Ross ([2010]) focus on some of the aspects of the methodology underlying many neuroeconomic studies and on the status of the evidence provided by these studies. On the basis of the studies they examine, they argue that currently the grounds for informing economic models with neural data are methodologically dubious and evidentially weak.

More recently, Fumagalli ([2011]) has put forward a novel critique. He claims that economists are provisionally justified in resisting the incorporation of 'neural insights' into their models of choice behaviour because of the tractability costs that such 'neural enrichment' is likely to impose on economic modelling.[1] In order to defend this claim, Fumagalli ([2011], p. 618) leverages economists' modelling practices, but also 'the pragmatic and epistemic goals which govern the construction and evaluation of models' in economics and in other sciences.

The arguments in support of his claim are twofold. The first argument is that building a descriptively accurate and neurally-informed model of choice behaviour entails that the model will contain many neural variables. Because such a model would contain many neural variables, it is likely that modellers will incur tractability costs that overweight the explanatory benefits yielded by the model. Insofar as modellers of choice behaviour are justified to incorporate neural variables in their models only if the tractability costs of this type of modelling are compensated by its explanatory benefits, modellers of choice behaviour should not build models that are neurally-informed.

The second argument begins from the premise that choice behaviour can be modelled at different levels (e.g. psychological, neural, biological and micro-physical). Then it claims that providing a descriptively accurate and tractable model of choice prevents economists from incorporating variables at the neural level. Since modelling choice behaviour at the neural level would involve too high modelling costs in comparison to models incorporating variables at some other level, economists should refrain from modelling choice behaviour at the neural level, and, more generally, from building models that span multiple levels, which would be 'prohibitively impractical.'

There are three reasons why considering Fumagalli's ([2011]) arguments is congenial to the main aim of the present paper, that is: to elucidate some of the core aspects of current modelling practice in neuroeconomics. First, these arguments provide a novel and up-to-date critique of neuroeconomics; second, being based on the notion of tractability, this critique purports to take actual methodology and modelling practice in neuroeconomics seriously; third, this critique is representative of a number of common misconceptions about current practice and goals of modelling in neuroeconomics. Two such misconceptions concern what it can take to neurally-inform an economic model, and what it means that neuroeconomic modelling spans multiple levels. Hence, Fumagalli's

([2011]) arguments can serve as a particularly useful foil for the issues explained in this paper.

By reconstructing a case-study from one prominent area of neuroeconomics, viz. Niv et al.'s ([2012]) work on risk-sensitive behaviour, I shall firstly explain in which sense there are already descriptively accurate, tractable economic models of choice that incorporate neural insights. In focusing on this case study, I shall clarify under which conditions it is plausible to say that a model is tractable. Finally, I shall explain how expressions such as 'neurally enriched economic models' or 'informing an economic model with neural insights' should be understood. A closer look at actual practice in neuroeconomics makes it clear that a neurally-informed model of choice need not literally include variables standing for properties of neurons such as having a certain refractory period or having a certain membrane resistance. To say that results from cognitive neuroscience can inform economic models of choice can mean at least two things: Neural results can, on the one hand, confirm (or disconfirm) the causal relevance of latent, subjective variables or processes posited by competing models of choice behaviour, and, on the other, point to causally relevant variables or processes overlooked by existing models.[2]

The paper is structured thus. Section 2 begins by reconstructing a case-study representative of model-based analyses of decision-making, which is a growing area in neuroeconomics (Corrado and Doya [2007]; Mars et al. [2010]; O'Doherty et al. [2007]). This case study will serve to elucidate what it can take to neurally-inform a model of choice. Drawing on this case, Section 3 argues that currently for at least some economic model of choice behaviour, the benefits derivable from neurally-informing an economic model do not involve special tractability costs (i.e. tractability costs that are too high and specifically due to neurally-informing a model of choice). This argument challenges whoever believes that currently 'a neural enrichment of economic models is likely to impose [significant tractability costs] on economists' (Fumagalli [2011], p. 627) to provide actual case-studies in support of their view. Section 4 explicates how expressions like 'neurally enriched economic model' are best understood. Section 5 concludes with a summary of the contribution of the paper to existing literature.

2 Neurally-Informed Models of Choice: A Case Study

This section reconstructs a representative study from a growing area in neuroeconomics: Niv et al. ([2012]). This case-study offers a solid basis for articulating the core claims of this paper.

First, this study poses a challenge to the claim that neuroeconomists have not shown that the modelling benefits of informing economic models of choice with neural insights compensate for their modelling costs (cf. Bernheim [2009]; Fumagalli [2011];

Harrison [2008]). Second, while the case-study illustrates model-building practice in neuroeconomics, it illustrates the claim often made in the economic as well as in the neuroscientific literature that the degree of descriptive accuracy of a model is jointly determined by the causal structure of the real-world system under investigation, the modeller's varying epistemic interests and purposes in relation to that system, and the modeller's audience (on 'descriptive accuracy' or realism in economics see e.g. Bhaskar et al. [1998]; Mäki [2007]). Finally, the case-study paves the way for explaining how tractability is better understood in (economic) modelling (Section 3) and for explicating how expressions like 'neurally enriched economic model' should be understood (Section 4).

It will be clear that the main benefit of the neuroeconomic approach in the case-study under consideration consists in a kind of independent test of competing models of risk-sensitive choice behaviour. The economic models are first fitted on the basis of choice data, and then compared to the neurobiological data whose patterns they can either fit or fail to explain. It will be clear that this added value of neurally-informed economic modelling does not need involve special tractability costs.

2.1 A case-study on risk-sensitive choice

Niv et al. ([2012]) set out to investigate the mechanism of humans' attitude to risk (defined as variance in the reward outcomes of their actions) during decision-making tasks where payoffs (or reward outcomes) must be learned from experience. Sensitivity to risk is a prominent economic phenomenon, which is explained in various ways by different economic models of choice behaviour. Niv and colleagues were interested in which one of a number of competing models best explained this phenomenon. Two traditional modelling approaches they considered were expected utility theory and reinforcement learning, which are widely used in economics.

Within expected utility theory, risk-sensitive behaviour is taken to be explained by some nonlinear subjective utility function that maps an agent's wealth (e.g. money) into his subjective utility for wealth (cf. Bernoulli [1954]; see Schoemaker [1982] for a critical review of this modelling approach to risk-aversion).[3] The general (or 'global') shape of a utility function implies a specific attitude towards risk. If the utility function is concave (e.g., the functional form of the utility function is logarithmic), then risk-averse behaviour is implied. If the utility function is convex (e.g., the functional form of the utility function is exponential), then risk-loving behaviour is implied. Hence, within expected utility theory, different models with different functional forms can be understood as hypotheses about how risk-aversion depends on wealth.

In standard reinforcement learning, risk-sensitive behaviour is taken to be explained by biases in the sampling of the available options due to the interaction between choice and learning (cf. March [1996]; Niv et al. [2002]). In standard reinforcement learning models, risk-sensitive behaviour does not arise because of some nonlinear utility function; instead, risk-aversion is ordered by the learning rate in the model, and higher learning rates imply more risk-aversion. Hence, within reinforcement learning, different models with different learning rates can be understood as hypotheses about how risk-aversion depends on learning dynamics.

2.1.1 Target and modelling framework

Niv et al. ([2012]) delimited their target to choice behaviour where there is no a priori knowledge about different payoffs and their probabilities. This type of choice behaviour requires that knowledge about payoffs and probabilities is acquired through experience. For investigating this type of target, Niv and colleagues turned to the field of reinforcement learning, which provides a modelling framework where decision-making can be quantitatively and computationally analysed (Sutton and Barto [1998]). Reinforcement learning is a branch of artificial intelligence and machine learning that offers a collection of algorithms to address the problem of learning what to do in the face of rewards and punishments received by taking different actions in an unfamiliar environment. Within reinforcement learning, agents are modelled as using past experience to estimate mean (expected) values of different options. Given a choice, agents choose between options based on their values.

Besides the nature of their target, there was a second reason why Niv et al. ([2012]) worked within this framework. Since the mid ‘90s of the last century, the reinforcement learning framework has become increasingly central to investigating the neural substrates of learning and decision-making (see Niv [2009] for a review). This centrality is justified by the discovery that the phasic firing of dopamine neurons in the midbrain substantia nigra pars compacta and the ventral tegmental area, recorded from primates engaged in reward-learning tasks, can be described as encoding a key learning signal (Montague et al. [1996]; Schultz et al. [1997]). This is the *temporal difference reward prediction error*, and it is used in several types of reinforcement learning algorithms (Sutton and Barto [1998], Ch. 6).

2.1.2 Research question and hypotheses

Niv and colleagues investigated a specific aspect of the mechanism of humans’ attitude to risk by asking whether and how risk sensitivity is integral to (i.e. is a component or constituent of) human reward-learning, and how it can best be modelled. If risk sensitivity is not integral to reward-learning, then there might be two separate mechanisms of risk-

sensitive choice. One might be the mechanism by which we learn mean values of different options—as per traditional reinforcement learning models. The other mechanism might enable us to learn the variance associated with options' payoffs (or reward outcomes). Instead, if risk sensitivity is integral to human reward-learning, then traditional reinforcement learning models may be extended to take account of this feature of human learning and decision-making.

2.1.3 Competitive models of risk-sensitive behaviour

Two ways were identified for extending standard reinforcement learning models so that they take account of risk sensitivity. First, some nonlinear subjective utility for different outcomes (as captured e.g. by Bernoulli's [1954]) could be incorporated in the standard temporal-difference model. Second, the model could explicitly distinguish the asymmetric effects that positive and negative rewards have on the learning process (as per Mihatsch and Neuneier [2002]).

Three models of risk-sensitive choice behaviour were considered within the framework of reinforcement learning. One was a standard reinforcement learning model. The other two were extensions of this model.

The core of the *standard temporal-difference* (TD) learning model is this update rule:

$$[1] \qquad V(S)_{new} = V(S)_{old} + \eta\, \delta(t_{outcome}),$$

where $V(S)$ denotes the value of a chosen option S, η is a learning rate parameter, and $\delta(t_{outcome})$ is the temporal difference reward prediction error computed at each of two consecutive time steps ($t_{stimulus}$ and $t_{outcome} = t_{stimulus} + 1$). The prediction error is the difference between experienced and expected reward outcome, computed as:

$$[2] \qquad \delta(t) = r(t) + V(t) - V(t - 1),$$

where $V(t)$ is the predicted value of some option at time t, and $r(t)$ is the reward outcome obtained at time t. The reward prediction error at $t_{outcome}$ is used to update $V(S)$, which is the value of the chosen option, according to [1]. The reward prediction error is thus used to update expectations in order to drive learning and guide decision-making. Although the standard TD-learning model does not explicitly take risk into account, it can allow for risk aversion due to biases in the sampling of the available options, as mentioned above.

The second model considered by Niv and colleagues' ([2012]) was what they called a *Bernoulli utility model*. As in the standard temporal difference model, the update rule of the utility model is [1]. However, this model differs from the first one for the prediction error that it is used to update $V(S)$:

$$[3] \qquad \delta(t) = U(r(t)) + V(t) - V(t - 1),$$

where $U(r(t))$ is the subjective utility of the reward outcome at time t. Because of this type of prediction error, the variance of reward outcomes is not taken explicitly into account in

this model either. For different reward outcomes, depending on the value of a utility parameter *a*, subjective utility curves could be either convex or concave, thereby implying respectively risk-averse and risk-loving behaviour.

The third model was a *risk-sensitive TD-model*, which penalizes or favours the variance of reward outcomes by distinguishing two update rules for positive and negative prediction error. These are:

[4'] $V(S)_{new} = V(S)_{old} + \eta^{+} \delta(t_{outcome})$, if $\delta(t_{outcome}) > 0$,

[4''] $V(S)_{new} = V(S)_{old} + \eta^{-} \delta(t_{outcome})$, if $\delta(t_{outcome}) < 0$,

so that if $\eta^{+} < \eta^{-}$, the effect of negative prediction error on learned values is larger than that of positive prediction error, leading to risk aversion, and vice versa if $\eta^{+} < \eta^{-}$.

For all three models, Niv et al. ([2012]) assumed a softmax action selection function, which yields the probability *Prob* (*A*) of choosing a stimulus (or option) *A* from estimated values of the stimulus (or option) *V*(*A*):

[5] $Prob\ (A) = \exp^{\tau V(A)} / (\exp^{\tau V(A)} + \exp^{\tau V(B)})$,

where τ is a parameter called inverse temperature. As τ tends to ∞, the stimulus with highest value has a much higher probability of being selected than the others. As τ tends to 0, all stimuli become equally probable.

Hence, the parameters involved in the three models are respectively: the learning rate η and the inverse temperature τ for the TD-model; the learning rate η, the inverse temperature τ, and the utility parameter *a* for the utility model; the learning rates η^{+} and η^{-} and the inverse temperature τ for the risk-sensitive TD-model.

2.1.4 Model-based fMRI: From economics to brains and back

By investigating their target within the reinforcement learning modelling framework, Niv and colleagues could integrate their competitive models directly into the design and analysis of a neuroimaging experiment. One benefit of testing economic models of decision-making in neuroimaging studies is that they allow patterns of brain activity to be related to variables in the models that may have no direct relationship with choice behaviour, as for the cases of risk sensitivity and learned values of stimuli. The identification of this relationship can often afford constraints to existing models of choice, with respect, for example, to the cognitive architecture of human information processing that a given model might presuppose (cf. Forstmann et al. [2011]).

Niv and colleagues scanned human participants using fMRI, while they made choices between two visual stimuli associated with equal mean monetary value but different variances. Participants had to learn about the payoffs associated with the visual stimuli so as to maximize their final monetary winning.

The three models could describe the transformations from the set of stimuli inputs presented to participants to choice behaviour in the task. In general, '[t]he specific

"internal" operations required to effect such a transformation are the variables of interest in the neuroimaging study, as it is these variables that will ultimately be correlated with the neuroimaging data' (O'Doherty et al. [2007], p. 36). In studies such as Niv and colleagues', model variables like the prediction error δ serve as proxies for internal, subjective decision variables, which are assumed to drive choice behaviour. These internal, subjective decision variables are used for searching for their corresponding patterns of brain activity in the neuroimaging data. Information about neural correlates of these internal subjective decision variables can inform alternative models of the same type of choice behaviour by helping identify its neurocomputational mechanism. Let me explain this point in reference to Niv et al.'s ([2012]) study.

Following a growing approach in model-based fMRI, Niv and colleagues firstly fit the free parameters of the three models to choices of each participant by using a maximum a posteriori estimation procedure. Then, in order to determine which model provided the best fit for each participant, they compared the posterior likelihoods of each participant's choice behaviour according to each model.

For the three models, there were two variables that changed on a trial by trial basis as a function of learning and decision-making: the learned values of the stimuli $V(S)$ and the prediction error δ. Once the best-fitting models parameters were found, the models were regressed against fMRI data for each participant to identify brain signals correlating with the model-generated prediction errors and values of the stimuli. The three models made qualitatively different predictions about the value signals for a particular chosen stimulus in the task. Relying on these prediction, Niv and colleagues extracted the 'neural values' of the different stimuli from fMRI BOLD signals. In this way, they could identify which of the three competing models was best supported by the neural data.

Specifically, Niv and colleagues entered the three competing models into a regression analysis against the fMRI data to determine which model provided a better fit to the fMRI data for a given brain region of interest. This brain region was located in the nucleus accumbens and was determined based on previous relevant research in neuroeconomics. Hence, by means of model-based fMRI, Niv et al. not only provided independent evidence about how a particular cognitive function might be implemented by certain patterns of neural activity, but, more important, they could also discriminate the explanatory power of competing economic models of choice behaviour.

It was found that both the traditional TD-learning model and the utility model were not supported by neural data. Instead, both behavioural and neural data supported the risk-sensitive TD-model, providing evidence that prediction error signalling in the nucleus accumbens is indeed sensitive to the variance of reward outcomes. These results suggest that any descriptively accurate model of human choice should take account of risk sensitivity, as per the risk sensitive TD-model.

2.2 Neurally-informed modelling

Niv et al.'s ([2012]) provides an example of how results from cognitive neuroscience can be used to inform economic models of choice. Such results can be used both to assess the explanatory power of existing models of choice as well as heuristically, to advance novel models. Neural results can, on the one hand, confirm (or disconfirm) the causal relevance of latent, subjective processes posited by competing models of choice behaviour, and, on the other, point to causally relevant processes overlooked by existing models. Does neurally-informing economic models in either of these ways involve special tractability costs?

With respect to this question, several critics of neuroeconomics share the concern that accurately modelling the neural substrates of choice behaviour most likely involves significant tractability costs because 'numerous cerebral systems are highly interconnected at the anatomical and physiological levels' and because 'several neural areas activate in a wide range of decision contexts' (Fumagalli [2011], p. 628 and p. 633; see also Bernheim ([2009]); Ortmann ([2008]); Vromen ([2010])). The concern is that in order to accurately 'neurally-inform' a model of choice behaviour one has to take account of the neural substrate of the choice-phenomenon of interest, which will implicate too high modelling costs because one would have to deal with a too complex system.

Even if we evaluate such criticisms at the same level of abstraction and generality, they do not provide telling reasons for believing that neurally-informed modelling of choice behaviour is likely to involve special tractability costs. To begin with, in light of actual scientific modelling practice and of researchers' specific epistemic goals, it should be obvious that there is no straightforward relationship between the number of variables (neural or non-neural) included in a model and the descriptive accuracy (or realism) of that model. The descriptive accuracy of a model is a function of the features included in the model that matter to the phenomena of interest displayed by the target. What matters and what doesn't is jointly determined by the causal structure of the real-world system under investigation, the modeller's varying epistemic interests and purposes in relation to that system, and the modeller's audience (cf. e.g. Craver [2009], pp. 590-1; Mäki [2012]).

Neuroeconomists do not pursue a generic ideal of descriptive accuracy for their models. The way in which Niv et al. ([2012]) used their models was aimed at understanding whether and how specific latent subjective decision variables are causally relevant to some phenomenon of economic interest displayed by the target modelled. The descriptive accuracy of the models that they considered should be assessed as a function of such purpose. 'The essential function of the model [in these types of studies] is not necessarily to serve as an explicit hypothesis for how the brain makes a decision, but only to formalize an intermediate decision variable' (Corrado and Doya [2007], p. 8180). In so

far as the variables of interest in the model stand for causally relevant features in the system, then the model is descriptively accurate. So, in so far as risk sensitivity is integral to reinforcement learning and the learning process is driven by nonlinear effects of unpredictable reward outcomes, risk-sensitive TD-models are descriptively accurate.

Secondly, the fact that the brain is a neural network does not entail that neuroeconomic models taking account of a particular type of neural variable *X* must also take account of other variables of types *Y* and *Z* just because neural circuits of types *Y* and *Z* are connected to circuits of types *X*. Instead of imposing special tractability costs, physiological and anatomical knowledge about the central nervous system can usefully constrain model-building in neuroeconomics because it can help to better delineate neural regions of particular interests and it can serve as heuristics for the search of a certain type of algorithms within a given neural architecture.

Anatomical constraints and previous physiological findings were in fact used by Niv et al. ([2012]) to justify their focus on a specific region of interest in the nucleus accumbens. This modelling choice was based on previous studies showing that reward prediction errors are correlated with activity in the nucleus accumbens of humans engaged in classical or instrumental conditioning tasks that involve monetary rewards. Moreover, anatomical criteria were used to better delineate the regions of interest within the nucleus accumbens for each participant.

Results about neural architecture can serve as a heuristic in the search for the mechanism of decision-making. Analogies have been drawn between the anatomy and connectivity pattern displayed by the basal ganglia, and certain neural architectures for implementing reinforcement learning algorithms (O'Doherty et al. [2007], p. 43). Specifically, actor-critic reinforcement learning algorithms have been proposed as routines implemented in specific regions of the basal ganglia partly because of the neural connectivity and anatomy of these regions (e.g., Joel et al. [2002]).

One further observation undercuts the objection that the anatomical and dynamic complexity of the human brain are most likely to make it intractable to accurately modelling the neural substrates of choice behaviour. This charge overlooks that structural constraints can be useful to infer the function performed by activity in some neural circuit, and hence to identify the algorithm that is likely to be implemented by activation in that circuit. The function of some particular neural circuit, that is, can be inferred from their pattern of connectivity with other structures (cf. Sporns [2011]).

Finally, from the fact that many neural circuits are involved in a wide range of decision contexts it does not follow that a) we cannot identify a restricted number of particular circuits differentially active in those decision contexts, and b) distinct decision contexts do not share some underlying property such that they can be thought of the same

type, as critics like Fumagalli ([2011], p. 628), Bernheim ([2009], p. 7) and Vromen ([2010], p. 180) appear to suggest.

If particular neural circuits are differentially active in certain classes of decision-contexts, then there is reason to believe that only a subset of the many neural variables involved in that type of decision-context is especially relevant to carrying out the cognitive functions underlying that decision. This is in fact what has been found in research on the neural bases of social decision-making. Altruistic, fair or trusting behaviour differentially activates reward-related brain areas such as the striatum and specific prefrontal circuits (see e.g. Lee [2008] for a review). So, there is reason to believe that some neural variables are especially relevant to identifying the mechanism of decisions carried out in certain types of contexts. Niv and colleagues relied exactly on this fact in the selection of relevant neural circuits on which their analysis focused.

Furthermore, at the level of Marr's ([1982]) computational analysis, a wide range of decision contexts are justifiably considered of the same type. This is the case of the range of decision problems addressed by reinforcement learning algorithms (Sutton and Barto [1998], Ch. 11). In spite of apparent differences, all these problems share underlying properties such that they are justifiably considered as different instances of the problem of learning what to do in the face of rewards and punishments received by taking different actions in an uncertain environment. If this is so, then the brain might use algorithms of the same family to solve seemingly different decision problems (cf. Dayan and Daw [2008]).

3 Tractability: When Does Size Matter?

What does it mean that a model is tractable? In general, it means that the model is easy to build, easy to analyse, or easy to manipulate. More specifically, in economics as well as in the philosophical literature about scientific and economic modelling, 'tractability' can denote a property of a *model itself* or a property of the activities of *modelling*.

Gabaix and Laibson ([2008], p. 294) characterise 'tractability' as a property of a model itself when they explain: 'Models with maximal tractability can be solved with analytic methods—i.e. paper and pencil calculations. At the other extreme, minimally tractable models cannot be solved even with a computer, since the necessary computations/simulations would take too long. For instance, optimization is typically not computationally feasible when there are dozens of continuous state variables—in such cases, numerical solution times are measured on the scale of years or centuries.' This understanding of tractability is widespread in economic modelling indeed.[4]

Less discussed in the economic literature is the sense of 'tractability' as a property of *modelling*. As explained by Hindriks ([2006]), 'tractability' in this sense picks out a property that comes in degrees, and that can evolve over history as a function of the modellers' cognitive, computational and material resources and of developments in science

and technology (on this sense of tractability cf. also Knuuttila and Loettgers [2012]; Mäki [2009]; Odenbaugh [2007]).

Before moving on to explicating for each of these two senses under what conditions a model is tractable, it is worth being clear on one point.[5] Let us stipulate that a model is tractable if it has some analytical solution. A solution to a model is 'analytical' if the model has a closed-form solution, in terms of a finite number of known functions. Consider a traditional expected utility model that has analytical solutions (e.g. the Bernoulli utility model). Suppose that this utility model fails to predict a set of relevant choice and neural data; some alternative, empirically-informed model that has no analytical solution successfully predicts the same set of data. This is a case where the loss in tractability of a model would trade-off with a gain in its predictive accuracy.

By itself, however, this loss in tractability cannot be used as a reason to prefer the more traditional, utility model. Tractability costs—at least understood in terms of lack of analytical solutions—can be used as an argument against neuroeconomics only if there are other significant advantages to using the traditional utility model. But that is precisely what is in question in comparing the predictive power of competitive models (cf. Harless and Camerer [1994] for an early statement and experimental exploration of this argument in economics). With this disclaimer in place, let us now better distinguish between tractability as a property (or a cluster of properties) of a model itself and tractability as a property (or cluster of properties) of the activities of model-building and model-use.

If we focus on the model itself, then mathematics and complexity theory offer an objective characterization of a tractable model (Garey and Johnson [1979]). According to this characterization, a model is tractable exactly if a Turing machine can provide an output for each input to the model within a certain number of steps. For a model to be tractable, this number of steps must be at most some polynomial function of the number of bits in the input. The tractability of a model itself concerns the time a Turing machine needs for implementing (or finding a solution to) the model. 'If the time needed cannot be described by a polynomial function of the length of the input, it tends to go to infinity. If this is the case, the [model] is intractable' (Hindriks [2006], p. 413).

In this sense, the three models considered by Niv et al's ([2012]) study were all tractable, given the structure and dimensionality of the action-space of the task the models had to address. There are a number of approaches to the 'curse of dimensionality' that RL-models such as the types of TD-learning models considered by Niv and colleagues face when they target more complex, high-dimensional systems (see e.g. Doya et al. [2002]; Wilson and Niv [2012]). So, provided suitable representations of relevant features of the task to be addressed, TD-learning models are generally tractable in the sense just discussed (Sutton and Barto [1998], Ch. 6).[6]

Two further points should be noted in light of the complex-theoretic characterisation of tractability. First, a definition of tractability of a model in terms of a function of its input length implies that the number of variables of a model per se is *not* a reliable indicator of its complexity, as claimed by Fumagalli ([2011]). The inherent complexity (or tractability) of a model is captured by *how* the demands on time and other computational resources increase with its input length when the model yields a solution, rather than by 'the fact *that* demand on computational resources increases with input size' (van Rooij [2008], p. 944). Second, since tractability, in this sense, is an inherent property of models themselves, the answer to the question 'Which of two models of a given target system is more tractable?' has an objective answer, which does not depend on modellers' interests and purposes or on current restrictions on human technology and theory.

Let us now consider the activities of model-building and model-use. If we understand tractability as a property (or cluster of properties) of modelling, rather than models themselves, then the tractability of a model can plausibly be identified with the ease with which modellers can build the model or manipulate it in order to obtain some desired result. The number of variables appearing in a model does seem to affect ease of model-building and model-use.

The main everyday reasons one wants to keep variable sets in specific models small are because of practical, material or cognitive limitations, which humans happen to have at a certain time in history due to limitations in current technology, theory or material resources. These are not deep issues of principle, but purely contingent, temporary constraints on modelling methodology, which do not apply only to neuroeconomics, but to scientific modelling more generally. Given this understanding of tractability in relation to modelling practices, the real issues for neuroeconomic methodology are two: *currently*, does informing economic models of choice with neural evidence require one to incorporate several variables in such models? *Currently*, are neurally-informed economic models of choice often, seldom or never easy to build or use? Section 2.2, above, already provided reasons for a negative answer to the first question. An adequate answer to the second question cannot be pursued at an unhelpfully abstract level, where modelling is considered in isolation from specific interests, purposes and capacities of actual modellers and of their audience, and in isolation from the type of causal structure of the system under investigation. If this is so, then that there may be no objective answer to the question of whether a neurally-informed model of risk-sensitive choice is easier to use or to build in comparison to an alternative neurally-*un*informed model.

One general consideration can defuse concerns about ease of model-building and model-use in at least certain approaches to neurally-informed modelling in neuroeconomics. Model-based fMRI studies in neuroeconomics such as Niv et al.'s ([2012]) are becoming more and more widespread partly because of the increasing level of

sophistication of the technology as well as of the mathematics, statistics and machine learning techniques available (cf. Daw [2011]; Friston and Dolan [2010]). In so far as the tractability of a model—in the sense under consideration—is relative to development in science and technology and to auxiliary theories in neighbouring fields (Hindriks [2006], pp. 413-4), model-building and model-use in areas of neuroeconomics such as model-based fMRI are becoming ever more tractable.

4 Neural Integration and the Co-evolutionary Research Ideology

What does it mean that neural variables should *inform* economic models of choice? Does the project of neuroeconomics consist in modelling human choice behaviour at one particular (lower-) level, viz. the neural level, rather than at some other level (cf. Fumagalli [2011]; Harrison [2008], Sec. 4.6; Kuorikoski and Ylikoski [2010])? And does the diversity of methodologies used by neuroeconomists implicate that neuroeconomics has no single unifying goal (cf. Ross [2008])?

Terms like 'incorporating,' 'importing,' and 'integrating neural insights,' or 'integrating neural variables' into economic models are often used in arguments concerning neuroeconomics without further qualification. Since they are generic, it is difficult to assess the plausibility of an argument such as Fumagalli's ([2011]) without knowing how they should be understood. I argue that expressions such as 'neural enrichment of economic models' are best understood in light of a co-evolutionary research ideology and within the familiar Marr's ([1982]) three-level of analysis explanatory framework, which are often advocated by researchers in the field of neuroeconomics, and by critics of neuroeconomics alike, to identify the ultimate goal of the neuroeconomics project (cf. Glimcher [2003]; Harrison [2008], Sec. 4).

There are various characterizations of what neuroeconomics is about (see e.g. Camerer [2008]; Glimcher [2009]; Rustichini [2009]). All these characterizations seem to agree on one fundamental point. In the words of Glimcher ([2009], p. 503): 'The goal of neuroeconomics is an algorithmic description of the human mechanism for choice.'[7] The appeal to an algorithmic description hints at a specific way in which neuroeconomics research spans multiple levels, involves insights from multiple disciplines, and can employ different methodologies.

In light of such a shared goal, current modelling practice in neuroeconomics is best understood within Marr's ([1982]) three-level of analysis framework. Marr's levels include the computational, the algorithmic and the level of implementation. The computational level specifies the problem to be solved in terms of some generic input-output mapping. In the case of neuroeconomics, this is the problem of selecting one option from a set of options available to the agent. Thus, the generic input-output mapping that defines the computational problem of neuroeconomics is a function mapping the set of options

available to the agent to an estimate of the agent's choice behaviour. It is generic in that it does not specify any class of rules for generating the output. This class is defined at the algorithmic level.

An algorithm specifies an effective procedure to solve a given problem. Neurally-informed models of choice behaviour belong to this level, as they aim to describe step-by-step procedures that produce an estimate of the agent's choice behaviour as a function of information about the choice options available. Hence, the models considered in Niv et al's ([2012]) study belong to the algorithmic level, as they specify alternative effective procedures for generating risk-sensitive choice behaviour. Accordingly, and in line with the core goal of neuroeconomics, Niv et al's ([2012]) goal is to advance our algorithmic understanding of choice behaviour.

The level of implementation is the level of physical parts and their organization. It describes the mechanism that carries out the algorithm. Mechanisms implementing an algorithm can in turn span different levels of organization. Obviously, the nervous system comprises different levels: neurotransmitters, neurons, populations of neurons, brain regions, and so forth. Different structures and activities at different levels of grain may be relevant to implement a given algorithm (Craver [2007], Ch. 5). The phasic activity of dopaminergic neurons in the basal ganglia, for example, probably encodes reward prediction errors. As mentioned above, the striatum, which is a region of the forebrain strongly innervated by dopaminergic fibers, might implement actor-critic RL-algorithms, which use reward prediction errors.

Once placed within Marr's three-level, asking whether 'human choice behaviour is more conveniently modelled at the neural—rather than some other—level' is a misunderstanding of neuroeconomics methodology (cf. Rustichini [2009]). Once it is clear that the relevant notion of a level involved in modelling in neuroeconomics is Marr's one, then it should be also clear that models at different levels do not compete with one another. To obtain a better understanding of choice behaviour, models are needed at each of the three levels. An economic model at the computational level targets the kinds of functions that can be optimized by a given type of behaviour. A neuroeconomic model at the algorithmic level targets the processes by which that behaviour can be carried out. A model at the implementation level targets the neural structures and activities that implement a given algorithm.

A *co-evolutionary research ideology* is congenial to the pursuit of an algorithmic understanding of choice behaviour and enables us to make sense of expressions like 'neurally inform' or 'neurally enrich' models of choice behaviour. This research ideology also justifies multiple methodologies for advancing our understanding of the computational mechanism of choice behaviour. According to Churchland ([1986]), *co-evolution* involves models put forth in one field of inquiry being susceptible to correction and

reconceptualization in light of discoveries, conceptual refinements, and methodologies made available outside the field at issue. There is no single, autonomous, privileged level of explanation. Rather, theories and models at one level should be constrained, revised and reshaped in terms theories and models at other levels (see also Colombo [2012]; Di Francesco, Motterlini and Colombo [2007]).

Consider claims like the following one: 'augmented economic models will also likely include results from sociology, anthropology, psychology, and other fields' (Park and Zak [2007], p. 54). One way to understand this claim is with Fumagalli ([2011], p. 629), when he comments: 'the point remains that simultaneously including constructs from various disciplines into a given economic model would often be prohibitively impractical.' But informing a given model with results and knowledge from other fields of inquiry does not mean that we should literally posit variables and constructs from various disciplines in our model. From the perspectives of Marr's three-level of analysis and of a co-evolutionary research ideology, informing a model with results and knowledge from various fields should rather be understood in the following way.

Consider models at each of Marr's three levels. Models at the computational level that identify which type of function may be optimized by some behaviour constrain possible algorithms. If a given algorithm is unable to solve the computational problems that our cognitive system needs to solve, then it cannot be the algorithm implemented by our nervous system. Or, if we find a deviation from a model at the computational level that predicts that agents deal optimally with noise (or uncertainty), depending on the aspect of the deviation, it might be concluded that noise has a different structure from what it was assumed, or alternatively that specific aspects of the algorithmic specification of the computational problem or of the underlying neural machinery affect behaviour in ways that make it sub-optimal. Taking these insights into account, a new computational model of how agents should deal with noise may be specified (see Beck et al. [2012] for a nice exploration of this type of case). Knowledge of the details of a putative neural implementation of an algorithm can render certain algorithmic models unfeasible. If, for example, the properties of a certain neural circuit are such that it cannot implement an algorithm with certain properties to solve a given problem, then the nervous system probably implements an algorithm with different properties or it solves a different type of computational problem altogether. So, informed by knowledge at the neural level of implementation, we can set out to build a new model. In light of these remarks, what is key to neurally-inform models of choice behaviour is 'a careful characterization of the deviation between models and experimental data' (Fernandes and Körding [2010], p. 346). This is evident in Niv et al.'s ([2012]) study as well as in Bayesian modelling in neuroscience (cf. Doya et al. [2007]).

Reinforcement learning and Bayesian modelling have been most powerful in pursuing such a co-evolutionary research ideology, so that models at the three levels inform each other in a fruitful way. Remarkably, for reinforcement learning and Bayesian cognitive neuroscience, the co-evolution of models and theories from different fields and at different Marr's levels has borne fruit not only for gleaning understanding about a wealth of phenomena related to perception, action, judgement and decision-making, but also for shaping the next generation of experimental research (cf. Niv [2009] on RL; on the project of Bayesian cognitive science see e.g. Tenenbaum et al. [2011]).

Working with Marr's distinction between levels, Niv et al.'s ([2012]) project illustrates co-evolution in practice in neuroeconomics. Their modelling project set out to advance our understanding of the computational mechanism through which risk may come to influence learning and decision-making. The normative framework of reinforcement learning was used to define the computational task solved by agents learning about, and choosing between options with the same mean (expected) value but different variances. Accordingly, learning was treated as aiming to select actions that will increase the probability of rewarding outcomes and decrease the probability of punishing ones. Choice behaviour was understood as aiming to optimize the consequences of actions in terms of a long-term measure of total obtained rewards. As already pointed out, within reinforcement learning, traditional models of choice behaviour do not explicitly represent the agent's attitude to risk, since the variance of different options is ignored.

If the specification of the computational task is revised as a function of findings at the algorithmic level, then co-evolution between computational and algorithmic models occurs. Among the algorithmic findings relevant to the computational level are the number of time-steps that each algorithm requires to compute the target function (cf. van Rooij [2008]), and the amount of behavioural variance explained by each model. Niv et al.'s ([2012]) finding that the risk-sensitive TD-model provided the best fit for participants' choice behaviour suggests that the computational-level model of our attitude to risk during reward-based learning and decision-making should be revised. The revised model at the computational level should take account not only of the mean reward values of options, as per traditional reinforcement learning models, but also such higher order moments as their variances.

As a co-evolutionary research ideology would prescribe, the models at the algorithmic level should be sensitive also to findings at the level of implementation. The three models considered by Niv and colleagues were integrated in the design of their fMRI experiment, thereby allowing for linking choice behaviour to neurophysiological and fMRI signals. Niv et al. ([2012]) scanned the level of implementation, looking for significant correlations between model-generated reward prediction error signals on the one hand, and participants' behaviour and brain activity on the other. By searching for relationships

between the level of implementation and the algorithmic level, that is, between neural variables and processes of interest and specific variables and processes in their algorithmic models, Niv and colleagues were able to provide evidence that risk sensitivity is integral to prediction error signalling in the nucleus accumbens. The evidence at the neural level suggested that algorithmic models of risk-sensitive learning and choice should apply a nonlinear transformation to prediction errors, rather than to outcomes (as in the utility model) to accurately model humans' attitudes to risk.

Furthermore, the co-evolutionary exchange apparent in Niv et al.'s study was also directed at advancing our understanding of the downstream effects of dopaminergic signals in the nucleus accumbens. From the finding that signals in the nucleus accumbens were associated with risk-sensitive learning, Niv and colleagues formulated a number of testable hypotheses about the neural mechanism for the asymmetric weighting of reward prediction errors, and about the broader neurocomputational architecture of risk-sensitive choice.

5 Conclusion

This paper has tried to show that, currently, for at least some economic model of choice behaviour, the benefits derivable from neurally-informing an economic model of choice do not involve special tractability costs. The claim challenges whoever believes that neurally-informing economic models is likely to impose substantial tractability costs on economists to provide actual case-studies in support of this belief. The paper has elucidated whether and in which sense a model's degree of accuracy and number of variables impact the model's tractability. More generally, the paper contributes to the debate on model-building at the interface between economics and neuroscience in two ways. First, it has explicated what it means 'to neurally-inform an economic model,' pointing to one of the added values of neuroeconomic modelling. Second, it has argued that modelling in neuroeconomics is best understood within David Marr's three-level of analysis framework and in light of a co-evolutionary research ideology.

Funding

Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" ([SPP 1516]).

Acknowledgements

I am sincerely grateful to Andrea Polonioli, Jan Sprenger, Liz Irvine, Michiru Nagatsu and Peter Dayan for helpful discussion, criticisms or suggestions on some of the ideas in the paper. Mauro Rossi was so kind to provide me with extensive feedback on previous versions of the paper. A special thank you goes to three anonymous referees for this journal for their time and their exceptionally constructive comments. The usual disclaimers about any error or mistake in the paper apply.

Matteo Colombo
Tilburg center for Logic and Philosophy of Science, Tilburg University,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
m.colombo@uvt.nl

References

Barberis, N., Huang, M., and Santos, T. [2001]: 'Prospect theory and asset prices', *Quarterly Journal of Economics*, **116** (1), pp. 1-53.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. and Pouget, A. [2012]: 'Not noisy, just wrong: the role of suboptimal inference in behavioral variability', *Neuron*, **74**, pp. 30-9.

Bernheim, B. D. [2009]: 'On the potential of neuroeconomics: a critical (but hopeful) appraisal', *American Economic Journal: Microeconomics*, **1**, pp. 1-41.

Bernoulli, D. [1954]: 'Exposition of a new theory on the measurement of risk', *Econometrica*, **22**, pp. 23-36.

Bhaskar, R., Archer, M., Collier, A., Lawson, T. and Norrie, A. (eds) [1998]: *Critical Realism*. London: Routledge.

Camerer, C. F. [2008]: 'The potential of neuroeconomics', *Economics and Philosophy*, **24**, pp. 369-79.

Camerer, C. F. and Loewenstein, G. [2004]: 'Behavioral Economics: Past, Present, Future', in C. F. Camerer, G. Loewenstein and M. Rabin, (eds) 2004 *Advances in Behavioral Economics*. Princeton, NJ: Princeton University Press and Russell Sage Foundation Press, pp. 3-51.

Camerer, C. F., Loewenstein, G. and Rabin, M. (eds) [2004]: *Advances in Behavioral Economics*. Princeton, NJ: Princeton University Press and Russell Sage Foundation Press.

Churchland, P. S. [1986]: *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, MA: MIT Press.

Colombo, M. [2012]: 'Constitutive relevance and the personal/subpersonal distinction', *Philosophical Psychology*. *iFirst* DOI:10.1080/09515089.2012.667623

Craver, C. F. [2007]: *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, C. F. [2009]: 'Mechanisms and natural kinds', *Philosophical Psychology*, **22**, pp. 575-94.

Craver, C. F. and Alexandrova, A. [2008]: 'No revolution necessary: neural mechanisms for economics', *Economics and Philosophy*, **24**, pp. 381-406.

Daw, N. D. [2011]: 'Trial-by-trial data analysis using computational models', in E. A. Phelps, T. W. Robbins and M. Delgado (eds), 2011, *Affect, Learning and Decision Making, Attention and Performance. Vol. XXIII*, Oxford: Oxford University Press, pp. 3-38.

Dayan, P. and Daw, N. D. [2008]: 'Decision theory, reinforcement learning, and the brain', *Cognitive, Affective & Behavioural Neuroscience*, **8**, pp. 429-53.

Di Francesco, M., Motterlini, M. and Colombo, M. [2007]: 'In search of the neurobiological basis of decision-making: Explanation, Reduction and Emergence', *Functional Neurology*, **22**, pp. 197-204.

Doya, K., Samejima, K., Katagiri, K. and M. Kawato, M. [2002]: 'Multiple model-based reinforcement learning', *Neural Computation*, **14**, pp. 1347-69.

Doya, K., Ishii, S., Pouget, A. and Rao R. P. N. (eds) [2007]: *Bayesian Brain: Probabilistic Approaches to Neural Coding*, Cambridge, MA: MIT Press.

Edmans, A. and Gabaix, X. [2011]: 'Tractability in incentive contracting', *Review of Financial Studies*, **24**, pp. 2865-94.

Fernandes, H.L. and Körding K. P. [2010]: 'In praise of "false" models and rich data', *Journal of Motor Behaviour*, **42**, pp. 343-9.

Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S. and Serences, J. T. [2011]: 'Reciprocal relations between cognitive neuroscience and cognitive models: Opposites attract?' *Trends in Cognitive Sciences*, **15**, pp. 272-9.

Friston, K. J. and Dolan R. J, [2010]: 'Computational and dynamic models in neuroimaging', *Neuroimage*, **52**, pp. 752-65.

Fumagalli, R. [2011]: 'On the neural enrichment of economic models: tractability, trade-offs and multiple levels of description', *Biology and Philosophy*, **26**, pp. 617-35.

Gabaix, X. and Laibson, D. [2008]: 'The seven properties of good models', in A. Caplin and A. Schotter (eds) 2008, *The foundations of positive and normative economics*. Oxford University Press, Oxford, pp 292–299.

Garey, M. R. and Johnson D. S. [1979]: *Computers and intractability: A guide to the theory of NP-completeness*. New York: Freeman.

Glimcher, P. W. [2009]: 'Choice: Towards a Standard Back-pocket Model', in P. W. Glimcher, C. F. Camerer, E. Fehr and R. A. Poldrack (eds), 2009, *Neuroeconomics: Decision Making and the Brain*, New York: Academic Press, pp. 503-21.

Glimcher, P. W. [2003]: *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Gul, F. and Pesendorfer, W. [2008]: 'The case for mindless economics', in A. Caplin, and A. Schotter (eds), 2008, *The foundations of positive and normative economics*, New York: Oxford University Press, pp. 1-40.

Harless, D.W. and Camerer, C. F. [1994]: 'The predictive utility of generalized expected utility theories', *Econometrica*, **62**, pp. 1251-89.

Harrison, G. [2008]: 'Neuroeconomics: a critical reconsideration', *Economics and Philosophy*, **24**, pp. 303-44.

Harrison G. and Ross D. [2010]: 'The methodologies of neuroeconomics', *Journal of Economic Methodology*, **17**, pp. 185-96.

Hindriks, F. A. [2006]: 'Tractability Assumptions and the Musgrave–Mäki Typology', *Journal of Economic Methodology*, **13**, pp. 401-23.

Hindriks, F. A. [2005]: 'Unobservability, tractability, and the battle of assumptions', *Journal of Economic Methodology*, **12**, pp. 383-406.

Joel, D., Niv, Y. and Ruppin, E. [2002]: 'Actor—critic models of the basal ganglia: new anatomical and computational perspectives', *Neural Networks*, **15**, pp. 535-47.

Knuuttila, T. and Loettgers, A. [2012]: 'The productive tension: Mechanisms vs. templates in modeling the phenomena', in P. Humphreys and C. Imbert (eds) 2012, *Representations, models, and simulations*, New York: Routledge, pp. 2-24.

Kuorikoski, J. and Ylikoski, P. [2010]: 'Explanatory relevance across disciplinary boundaries: the case of neuroeconomics', *Journal of Economic Methodology*, **17**, pp. 219-28.

Lee, D. [2008]: 'Game theory and neural basis of social decision making', *Nature Neuroscience*, **11**, pp. 404-9.

Mäki, U. [2007]: *Realism and Economic Methodology*. London: Routledge.

Mäki, U. [2012]: 'The truth of false idealizations in modelling', in P. Humphreys and C. Imbert (eds), 2012, *Models, Simulations and Representations*, New York: Routledge, pp. 216-33.

Mäki, U. [2009]: 'Realistic realism about unrealistic models', in H. Kincaid and D. Ross (eds) 2009, *Oxford Handbook of the Philosophy of Economics*, Oxford: Oxford University Press, pp. 68-98.

March, J. G. [1996]: 'Learning to be risk averse', *Psychological Review*, **103**, pp. 309-19.

Marr, D. [1982]: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Mihatsch O. and Neuneier, R. [2002]: 'Risk-sensitive reinforcement learning', *Machine Learning*, **49**, pp. 267-90.

Montague, P. R., Dayan, P. and Sejnowski, T. J. [1996]: 'A framework for mesencephalic dopamine systems based on predictive hebbian learning', *Journal of Neuroscience*, **16**, pp. 1936-47.

Morgan, M. S. and Knuuttila, T. [2012]: 'Models and modelling in economics', in: U. Mäki (ed) 2012, *Philosophy of economics. Handbook of the philosophy of science*. Amsterdam: Elsevier, pp. 49-87.

Niv, Y. [2009]: 'Reinforcement learning in the brain', *Journal of Mathematical Psychology*, **53**, pp. 139-54.

Niv, Y., Edlund, J. A., Dayan, P. and O'Doherty, J. P. [2012]: 'Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain', *Journal of Neuroscience*, **32**, pp. 551-62.

Niv, Y., Joel, D., Meilijson, I. and Ruppin E. [2002]: 'Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors', *Adaptive Behavior*, **10**, pp. 5-24.

Odenbaugh, J. [2003]: 'Complex systems, trade-offs and mathematical modeling: Richard Levins' "Strategy of model building in population biology" revisited', *Philosophy of Science*, **70**, pp. 1496-507.

Ortmann, A. [2008]: 'Prospecting neuroeconomics', *Economics and Philosophy*, **24**, pp. 431-48.

Park, J. W. and Zak, P. J. [2007]: 'Neuroeconomic studies', *Analyse & Kritik*, **29**, pp. 47-59.

van Rooij, I. [2008]: 'The tractable cognition thesis', *Cognitive Science*, **32**, pp. 939-84.

Ross, D. [2008]: 'Two styles of neuroeconomics', *Economics and Philosophy*, **24**, pp. 473-84.

Rubinstein, A. [2008]: 'Comments on neuroeconomics', *Economics and Philosophy*, 24, pp. 485-94.

Rustichini, A. [2009]: 'Is there a method of neuroeconomics?', *American Economic Journal: Microeconomics*, **1**, pp. 48-9.

Schoemaker, P. [1982]: 'The expected utility model: Its variants, purposes, evidence and limitations', *Journal of Economic Literature*, **20**, pp. 529-63.

Schultz, W., Dayan, P. and Montague, P. R. [1997]: 'A neural substrate of prediction and reward', *Science*, **275**, pp. 1593–9.

Sporns, O. [2011]: *Networks of the brain.* Cambridge, MA: MIT Press.

Stigler, G. J. [1950]: 'The development of utility theory II', *The Journal of Political Economy*, **58**, pp. 373-96.

Sutton, R. S. and Andrew G. Barto, A. G. [1998]: *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. and Goodman, N. D. [2011]: 'How to grow a mind: statistics, structure and abstraction', *Science*, **331**, pp. 1279-85.

Vromen, J. [2010]: 'Where economics and neuroscience might meet', *Journal of Economic Methodology*, **17**, pp.171-83.

Wilson, R. C., Niv, Y. [2012]: 'Inferring relevance in a changing world', *Frontiers in Human Neuroscience*, **5**:189.

[1] There is some ambiguity in what Fumagalli ([2011]) is actually claiming. At times he claims that 'a neural enrichment of economic models *is likely* to impose [tractability costs] on economists' (p. 627). Other times, he makes the stronger, distinct claim that 'elaborating descriptively accurate NE [ie neuroeconomic] models of decision making would *require* them [ie economists] to build *rather* intractable representations' (Ibid., emphases added). The argument put forward in the present paper challenges the more plausible, weak claim.

[2] Psychological results may inform economic models in similar ways, as much work in behavioural economics indicates (Camerer, Loewenstein and Rabin [2004]). Psychological data about e.g. memory, attention, learning, emotion, motivation, and personality traits may be correlated with the latent, subjective variables or processes posited by a model of choice behaviour. These correlations can become a source of evidence about the psychological plausibility of a target latent decision variable, about its causal relevance and about its possible causal relationships with other variables that may be overlooked by the model. This should not suggest that neuroeconomics brings no additional evidential or explanatory pay-offs. Rather, it suggests that the types of arguments put forward in this paper might carry over to psychologically-informed economic modelling. As it will be explained in Section 4, concepts, results and evidence from behavioural economics,

psychology and neuroscience can fruitfully be used to pursue a co-evolutionary research program aiming at advancing our algorithmic understanding of human choice behaviour.

[3] In economics as well as in philosophy of economics, there is little agreement on what precisely constitutes an adequate explanation of some behavioural regularity or economic phenomenon of interest. For the purposes of this paper, suffices it to clarify that to say that a utility function 'explains' some behavioural regularity or some economic phenomenon of interest means *at least* that the function fits reasonably well data relevant to that behavioural regularity or economic phenomenon. This clarification was prompted by an anonymous referee, for which I am grateful.

[4] See e.g. Stigler ([1950]), Barberis, Huang and Santos, T. ([2001]), Camerer and Loewenstein ([2004]), Edmans and Gabaix ([2008]). For philosophical discussions of this notion, see Hindriks ([2005], [2006]), Mäki ([2009], Sec. 9), and Morgan and Knuuttila ([2012]).

[5] This consideration was suggested to me by one of the referees, for which I am grateful.

[6] As one referee made me notice, one limitation should be pointed out of Niv et al.'s evidence, and more generally of most results in neuroeconomics. Niv et al. ([2012]) focused on what might be called 'local' risk aversion, i.e. stochastic dominance of preference for surer outcomes on *short timescales*. Economists are typically interested in attitudes to risk over *indefinitely extended timescales*. Given the types of instrumental and classical conditioning tasks used in their experiment, the Niv et al.'s findings might be relevant only to relatively stimulus- bound, 'local risk.' Economic models of risk attitude are generally applied to more stimulus- independent and non-local risk (e.g., in asset markets, risk of long- run inflation). So, the Niv et al. findings may be relevant to economic models of risk attitude at some scales of choice but not others. It is noteworthy, moreover, that even if the tractability of a model is scale-relative, empirically-demonstrated heterogeneity of risk attitudes at extended timescales might well be partly explained by people's varying levels of short-timescale, striatally-coded risk sensitivity, which only studies such as Niv et al.'s ([2012]) can confirm and measure.

[7] Whether the achievement of this goal will have revolutionary consequences for economics is a separate issue, which goes beyond the scope of this paper (on this issue see Craver and Alexandrova [2008]).