Model-Based Theorising in Cognitive Neuroscience

Abstract:

Weisberg (2006) and Godfrey-Smith (2006, 2009) distinguish between two forms of theorising: data-driven ‘abstract direct representation’ and modeling. The key difference is that when using a data-driven approach, theories are intended to represent specific phenomena, so directly represent them, while models may not be intended to represent anything, so represent targets indirectly, if at all. The aim here is to compare and analyse these practices, in order to outline an account of model-based theorising that involves direct representational relationships. This is based on the way that computational templates Humphreys (2002, 2004) are now used in cognitive neuroscience, and draws on the dynamic and tentative process of any kind of theory construction, and the idea of partial, purpose-relative representation.

1: Introduction

Weisberg (2006) and Godfrey-Smith (2006, 2009) distinguish between two forms of theorising: data driven abstract direct representation (ADR) and modeling. The key difference is that when theorising according to ADR, the target of investigation is the phenomenon (or more accurately, the model of data) that the theory is about. This means that the theory directly represents the phenomenon. In contrast, in model-based theorising, the model is the target of investigation, which is only later (if at all) compared to data about specific phenomena. Here, if a model represents a phenomenon, it does so only indirectly.

The aim of this article is to identify the specific kind of theorising practices that Weisberg and Godfrey-Smith draw on in order to make the distinction between direct and indirect representation, and in doing so open up a space for an alternative kind of model-based theorising. These alternative practices are found in cognitive neuroscience (and likely elsewhere) and involve models that directly represent their targets. Drawing on the work of Humphreys (2002, 2004) and Knuuttila and Loettgers (2011) it is suggested that the kind of modeling practices described by Weisberg and Godfrey-Smith are most clearly realised in the construction and analysis of (uninterpreted) computational templates. Other practices aim at modeling specific target phenomena and work in a rather different way. Weisberg does discuss target-directed modeling (2010, 2013 esp. pp. 74-97), but not explicitly in the context of theorising, so the article should be seen as an extension and critique of this work.

Offered here is a description of a non-trivial, distinct, and fairly common form of modeling; that of using computational templates to theorise about (classes of) phenomena that, given the current state of relevant scientific research, would be hard to theorise about in any other way, for example using purely data-driven ADR. It is argued that this form of theorising includes direct representation of target phenomena. This is done by minimising the differences between modeling and ADR, and in emphasising the nature of scientific representation as partial, purpose-relative, and crucially, as often tentative and in need of empirical testing[1].

Specifically, it is argued that modeling in cognitive neuroscience proceeds by scaffolding relevant background empirical and theoretical knowledge about a target system, including knowledge about its robust properties, into a theoretical structure using appropriate translations of computational templates. Plausibility constraints are used to try to ensure that relevant parts of these models and modelling frameworks are representationally adequate for their purposes. Theoretical progress is made by giving further justification and details to the claim that (parts of) the model represent (parts of) cognitive systems for specific purposes. These models are then used to generate theoretical inferences about the target system. In this form of model-based science, models represent their targets directly, if at all, even if this is only in a very abstract way, and very partially, in initial stages of model development.

The structure of the paper is as follows. Section 2 introduces and contrasts modeling and ADR practices and analyses how they fit into a broader landscape of aims and practices of theorising. Section 3 outlines the steps of model-based theorising in cognitive neuroscience and details the role of computational templates and plausibility constraints in modelling. The representational status of these models is discussed in Section 4, and illustrated with an example in Section 5. This example also shows how the notion of robustness comes into the early stages of model construction and the assessment of a model's basic representational capacities.

Section 2: Modeling and ADR

Weisberg (2007) and Godfrey-Smith (2006, 2009) set up the distinction between the practices of model-based theorising (indirect representation) and data-driven theorising (abstract direct representation) according to differences in what scientists actually do: "The practices are distinguished by the actions and intentions of theorists, not by the outcome of the process of theorizing" (p. 222, Weisberg 2007).

In abstract direct representation (ADR), theorists abstract away from (models of) data to form representations. This can include organizing and relating data, postulating causal relationships, and so on. Here, once the target is adequately represented, to analyse the representation is to analyse the target, as the representational relation is a direct one. In contrast, theorizing by modeling proceeds by constructing a model, and studying the model as an 'autonomous object' (Weisberg 2007, p. 224). Once the model is analysed, it may or may not be compared to target phenomena to see if and where it can apply. Here, when the model does represent, it does so only indirectly. So, it is the difference between the intentions of researchers during the stage of analysis that is supposed to be important – whether or not researchers in fact treat the theoretical structure as actually representing anything.

The steps associated with these kinds of theorizing are described below, and further similarities and differences between them are discussed. These are used to illustrate some ambiguities in this discussion between different types of theorizing, and

[1] The notion of representation is bracketed here, but it can be read in terms of Giere's (1988) notion of similarity (also discussed in Weisberg 2007, p. 217).

different types of representational practices. This opens space for other kinds of model-based theorizing, including the kind argued to be used in cognitive neuroscience.

2.1: Abstract Direct Representation (Data-driven)

Step 1: Abstract from phenomenon

In the ADR practice, a researcher first abstracts away from features of a target phenomenon, identifying more or less relevant properties and their relations, and constructs a representation of the target. In the cases described by Weisberg and Godfrey-Smith, this more often takes the form of organising and relating different groups of data. So, Mendeleev identified the properties of chemical elements that allowed him to organise them in a periodic table (Weisberg 2007, pp. 212-216), and Darwin used observations of atolls (circular islands containing a lagoon) in the Pacific and geological knowledge to theorise about the origins of these formations (Weisberg 2007, p. 227). In Godfrey-Smith's example, David Buss drew on selectionist principles and research on relations between reproduction at the cellular and whole-organism levels to theorise about how evolutionary transitions occurred (the creation of new evolutionary 'units', from single cells, to multi-cellular life, to organisms like us).

Step 2: Analyse representation

Next, the researcher analyses her representation, for example to generate predictions or identify ways of intervening on a system. So, Mendeleev could give explanations about the reactivity of certain elements based on their atomic structure, Darwin could make predictions about where atolls would develop, and Buss could make claims about the conditions likely required for the evolution of different types of life forms. As theoretical representations are intended to be representations of the target phenomenon, analysis of the representation counts as (direct) analysis of the phenomenon: "Because the theorist is analyzing a representation that is directly related to a real phenomenon, anything she discovers in her analysis of the representation is a discovery about the phenomenon itself, assuming that it was represented properly" (Weisberg 2007, p. 227).

2.2: Model-Based Theorising

Step 1: Construct a model

Weisberg (2007) illustrates the first step in modeling with the example of Volterra's construction of a model of predator and prey populations to study fish populations in the Adriatic (see e.g. Roughgarden 1979). Weisberg (2007) notes that Volterra "did not arrive at these model populations by abstracting away properties of real fish" (p. 210), and "did not start by looking for patterns in data" (p. 222). Instead, he (somehow) came up with a model that contained a few variables, related to each other in a fairly simple way. These included sizes of prey/predator population, growth and reproduction rates. Likewise, according to Godfrey-Smith (2006), Maynard Smith and Szathmáry used a modeling approach to theorise about evolutionary transitions by creating "idealized, schematic causal mechanism[s]" (p. 732). The idea is that they in some way selected appropriate simplifications and abstractions of possible classes of target phenomena without taking much note of biological research on real organisms.

Step 2: Analyse the model

Having constructed the model, the modeler then analyses it as an 'autonomous object' (Weisberg 2007, p. 224). Depending on the type of model constructed, this could involve inspecting or experimenting with it (e.g. in the case of scale models), solving sets of equations analytically, or by running a suitable simulation (Winsberg 2010). A range of things can be established at this stage. The model could show that a phenomenon can occur given a range of different (perhaps contradictory) assumptions about its underlying causal structure, thus undermining necessity or impossibility claims (Schlimm 2009). A model could also show why certain phenomena are rarely or never found in the natural world by exploring the dynamics of a world where they do exist.

Perhaps more often, analyses of models tell us about the general principles underlying a set of similar phenomena. Volterra analysed his model to see what kinds of general dynamic principles it generated, and found the surprising result (named the Volterra Principle) that when both predator and prey populations are decreased, this leads to an increase in the relative number of prey. Maynard Smith and Szathmáry explore their models to see how one kind of formally described life-form could evolve into another formally described life form. Accordingly, Godfrey-Smith (2006) notes that Maynard Smith and Szathmáry's claims, based on modeling, have a wider modal 'reach' than those of Buss, who theorized according to ADR, because "if they work at all, [they] would work just as well in a range of nearby possible worlds that happen to be inhabited by different organisms" (p. 732).

Step 3: Relate the model to target system

This stage is not always required (e.g. when models are used to undermine necessity claims). It is 'left to the reader' to complete in some disciplines (e.g. see Sugden on minimal economic models 2000, 2009), or can be done more or less explicitly by the modeler. The notion of 'construal' discussed by Godfrey-Smith (2006, 2009) and Weisberg (2007), provides a general way of thinking about the different ways that models can represent target systems, depending on the aims of the modeller. According to Weisberg (see esp. pp. 219-221), a researcher can construe a model according to four major criteria: identifying which parts of the model are relevant for the target (assignment), which parts of the target are represented in the model (scope), how well the model predicts the behaviour of the system and how well it describes the internal structure of the system (dynamical and representational fidelity criteria respectively). In this way, researchers can 'construe' the same model for a range of different representational purposes.

Crucially for the accounts of model-based science proposed by Godfrey-Smith and Weisberg, addressing the representational capacities of a model is the final stage of modeling. Only after constructing and analyzing the model do researchers explore whether the model can in fact be used to represent (relevant parts of) target systems at all. This is what makes models only indirectly represent target systems, as the intentions of the researcher initially only focus on exploring the model itself.

2.3: Similarities and Differences

The primary difference between ADR and modeling that Weisberg and Godfrey-Smith emphasise is between the intentions and practices of researchers. Modelers in a sense 'forget' about real phenomena while they construct and analyse their models, so need to add an extra step of relating the model to target phenomena (Step 3) in order to learn about target phenomena. Those engaged in ADR do not have this problem, as their theorising is aimed at analysing the target phenomenon at all stages. However, bringing out further similarities and differences between these ways of theorising highlights some misleading assumptions and false contrasts in Weisberg and Godfrey-Smith's discussion, and so opens up a wider range of modeling practices.

Similarities

In both modelling and ADR, the key similarity is that both involve abstraction and idealisation. In ADR, researchers abstract from (models of) data in a rather direct way. Modelers also abstract, though Weisberg and Godfrey-Smith say rather less about how they do so. There are at least two possible types of abstraction in modeling practice that Weisberg and Godfrey-Smith leave as ambiguous.

First, modelers can construct models of (classes of) phenomena (e.g. predator-prey systems), by abstracting away from examples of target phenomena (e.g. sharks and fish). Here, all the terms in the model are related to features of the (class of) target phenomena and are labelled as such (e.g. numbers of prey, death rate of predators). Of course, some features of the target phenomena may be left out, as with ADR. In this case the kind of abstraction present in the model may in fact be fairly similar in type to that found in ADR, but perhaps expressed more formally in the model. For example, it seems that researchers engaged with ADR may well also come up with the same abstract generalisations as Volterra, but based on observations from multiple experimental predator-prey systems.

In contrast, another kind of model-based theorising that Weisberg and Godfrey-Smith might be hinting towards are cases where modelers construct highly abstract, highly general mathematical models, that are abstracted from stylised, generic target phenomena, and where the terms in the models are left entirely uninterpreted. Here variables are just mathematical terms, and not, for example, 'number of predators'. These models really can be analysed as objects independently of any representational relations they may have with real phenomena, and a translation step (step 3 of model-based theorising) is required in order to see them as representational at all.

Indeed, Knuuttila and Loettgers (2011) argue that both types of practice can be found respectively in the work of Volterra and Lotka, who both arrived at the same equations, but using very different methods. Volterra took the first route of creating an abstract model of a specific phenomenon with the intention of answering specific questions about the causal dynamics underlying changes in fish and shark populations. Lotka on the other hand took the second route of constructing a highly general, uninterpreted template, which he then applied to a range of phenomena, including biological and chemical cases. Accordingly, Knuuttila and Loettgers (2011) write:

"…whereas Volterra approached modeling from the perspective of the causal explanation of real mechanisms, Lotka approached it from the perspective of applying a general template to specific cases." (p. 17)

In this case, looking at the aims behind abstraction and idealisation found in ADR and modeling can help distinguish between model-based practices that are not differentiated by Weisberg and Godfrey-Smith. First, models can be constructed using 'low strength' abstractions and idealisations from a few target systems, similarly to ADR practices, where terms are interpreted and easily applied to a small class of phenomena. In contrast, other models can be constructed using 'high-strength' abstractions and idealisations from stylised, generic phenomena, to form uninterpreted templates that can be applied to a much wider range of phenomena. Looking to other ways that model-based theorising is characterised by these authors helps to further identify the kind of practices they describe within a broader space of model-based practices.

Differences

As noted above, the main difference between the two types of theorising is supposed to be in terms of the degree to which the model or representation is analysed independently of analysing the target phenomenon. However, other differences are hinted at, particularly by Weisberg. Again, the distinction between models and templates is useful in identifying what is going on.

First, model-based theorising is sometimes aligned with exploratory theoretical work. So, Weisberg (2007) writes: "…[Volterra's] model suggested that the shark population would be especially prosperous. This is not something that Volterra or anyone else would have expected *a priori*" (p. 211, see also Weisberg 2013 pp. 88-89). The idea is presumably something like this: model-based theorizing is about taking a simple system and seeing what happens in it. That the mechanism behind increases in the shark population is surprising, is taken to show that the work really is exploratory, rather than simply trying to read patterns off data[2].

However, this presents something like a false contrast. An attempt to understand population dynamics using ADR would also have provided the surprising result that stopping fishing dramatically increases the number of shark compared to fish. That is, the surprisingly result can be identified using ADR and controlled experimentation as well as with modeling. Further, making non '*a priori*' predictions or surprising causal claims is surely an aim of any kind of theorizing. It does not seem specific to modeling itself that researchers can abstract (from somewhere) in order to theorise about the possible underlying mechanisms of a system, and use this to make novel predictions. Some modeling work clearly is highly exploratory in the way hinted at by Weisberg (play around with simple abstract systems and see what happens), but it need not be, and this need not be strongly tied with the ability to uncover surprisingly theoretical knowledge.

[2] Though a reviewer pointed out that the result was not in fact surprising: Darwin had already written about it.

Second, it is stated that ADR more usually aims at completeness: “Theorists who practice ADR typically aim to give complete representations” (Weisberg 2007, p. 229). Those engaged in modeling are supposedly happy with much simpler representations. In some cases this is true. Sometimes modelers might be interested in highly general principles that govern a range of systems, while those engaged with ADR might be more interested in uncovering details of specific mechanisms.

However, there are two reasons when this might fail. One is related to the type of target phenomenon. For distinct or unique target phenomena whose behavior is highly fragile, modeling as well as ADR may aim at completeness, as this is the only way to really explain the phenomenon, but both can be used as complementary ways of learning about a system. A second related factor (discussed further below) is how much data is available about a target system. Modeling may be a more viable way of gaining theoretical knowledge or structuring research in early stages of research programs, when there is simply not enough data around to get ADR off the ground. Here, modeling may ultimately aim at completeness, but not achieve it in the meantime due to very real constraints on the amount of data available (see e.g. Scholl and Räz 2013 on the development of Volterra’s model).

Third, according to Weisberg and Godfrey-Smith, in modeling what one finds out during the stage of analysis is “distinct from, and usually more general than, the system which inspired it” (p. 223, Weisberg 2007). So Volterra, whatever his aims, identified a very general relationship that applies to a wide range of predator-prey systems, not just fish and sharks. In contrast, Mendeleev working under ADR, and trying to find a way to organize elements and their properties, ‘only’ found out about basic chemical properties. Here there is a contrast between generating only local knowledge of about the target system (Mendeleev) and generating general principles that apply to many more systems than the one first considered (Volterra).

Here the distinction between models and templates (uninterpreted mathematical frameworks) becomes useful again. As noted earlier, Godfrey-Smith stated that Maynard Smith and Szathmáry’s claims about evolutionary transitions, based on modeling, have a wider modal ‘reach’ because they can apply to any possible worlds where the models have some (plausible) interpretation, not just to the case of biological evolution on Earth. However, the models developed by Maynard Smith and Szathmáry still only apply to evolutionary transitions, not to other phenomena. Templates however have broader application still. Lotka’s approach of treating the (Lotka-Volterra) equations as a template also made it possible to apply them to describe chemical, as well as biological, systems, (see e.g. Lotka 1920 on application in chemistry; Goodwin 1963 and Loettgers 2007 on application to circadian rhythms; Knuuttila and Loettgers 2011 and Weisberg 2013, pp. 77-78 for general discussion). That is, the theoretical structures that Maynard Smith and Szathmáry identify apply across a broader range of contingent factors than those of Buss, but still only to the same class of target phenomena (evolutionary transitions). The products of the analysis of templates can be distinct and general in a more spectacular way.

This discussion illustrates that to get a better grip on the landscape of model-based theorizing, it is a good idea to separate out different kinds of systems, and different theoretical approaches taken towards them. Modeling approaches can be exploratory, but they need not be, and a valuable property of any kind of theorizing is the ability to generate novel predictions. ADR may aim to generate more complete representations as compared to modeling, but not necessarily when the target system is highly fragile or unique, and in either case there is a serious constraint on the achievement of completeness in terms of the range of data available. Modeling may also aim to generate more general theoretical principles, but care needs to be taken to differentiate between the aims of constructing models of specific phenomena (local knowledge), classes of similar phenomena (more general knowledge), and the aims of constructing computational templates (potentially very broad application).

The rest of the paper is aimed at developing some of these distinctions in order to outline the kind of model-based theorizing found in cognitive neuroscience. Below is a discussion of the two key features of model-based theorizing in cognitive neuroscience, and an analysis of typical steps of model construction and development. This is followed by a discussion of the kind of representational relations found in this kind of model-based practice: direct, partial, and often under investigation. An example helps to further elucidate the ways that models represent their targets, which rests on an early epistemic role of knowledge of robust properties of templates and targets.

3: Steps of Model-Based Theorising in Cognitive Neuroscience

Modeling is prevalent in the cognitive sciences, perhaps unsurprisingly so. In cognitive neuroscience in particular, the complexity of the target systems under investigation, combined with the fact that experiments are often hard and expensive to run (experiments are often impressive technological feats in themselves), means that modeling and simulation is often the only viable way of making theoretical headway. Modeling, using computational templates, is used to scaffold meagre data into theoretical frameworks. This may initially sound as though any representation that occurs can only be indirect, but the following sections show how research methods used in cognitive neuroscience fit better with a story of direct representation. They show how the difference between the development of theoretical structures in ADR and modeling can be fairly minimal, and build on the idea of the development and testing of purpose relative, partial representations.

3.1: The Key Ideas: Templates and Plausibility

In cognitive neuroscience models are often developed from computational templates (Humphreys 2002, 2004, also Winsberg 2010). Templates are abstract syntactic schemata, such as sets of equations or computational methods that are interpreted to generate models of particular phenomena. In cognitive neuroscience, these can include equations from machine learning, such as reinforcement learning, (e.g. Corrado and Doya 2007, Glimcher 2011), sequential analysis and Bayesian analysis from statistics (Gold and Shadlen 2007, Griffiths et al. 2010), models of expected utility from economics (Glimcher et al. 2005, Sanfey et al. 2006), and ideas from thermodynamics used to model resource costs in cognitive processing (Ortega and Braun 2013, Friston 2005).

Humphreys outlines how templates are (mathematically, statistically) constructed, but of more interest here is how they are applied in particular cases. In order to use a template to construct a model, a ‘correction set’ is generated that de-idealises and deabstracts parts of the template, and changes the constraints and approximations of the template. The correction set is based on relevant background empirical, theoretical and practical knowledge, and in generating the correction set, modelers become aware of which parts of the model are likely to represent relevant parts of a target phenomenon for a specific aim (Humphreys 2002, pp. 76-81).

Weisberg (2007) comments that constructing models using ‘off-the-shelf’ templates is one way of proceeding in model-based science, but this is relegated to a footnote. He states: “In less path-breaking investigations, modelers often use ‘off-the-shelf’ models, structures that have already been applied to other phenomena of interest. In such cases, the first stage of modeling involves identifying the appropriate model, rather than explicitly constructing it” (Weisberg 2007, p. 222, footnote 8). Similarly, in Weisberg (2013) there is some discussion of templates (see e.g. pp. 75-76), but they are not named or discussed explicitly, or clearly differentiated from general but interpreted models, such as Volterra’s equations of predator-prey systems vs. Lotka’s computational template.

In fact, it is not always the case that the use of computational templates is related to ‘less path-breaking’ research, which presumably means that it provides less exciting or novel theoretical ideas. Suggesting that one system has the same underlying structure as another system can often be a major theoretical advance. For example, the use of reinforcement learning models (Bayer and Glimcher 2005, Schultz et al. 1997), originally from machine learning, as models of dopaminergic activity during decision making was a major theoretical advance that is now spurring much new research (see e.g. Glimcher 2011; Niv et al 2012 for contemporary work).

Further, using and interpreting templates may reflect a different motivation within modeling. For example, Knuuttila (2011) claims that modeling is often driven by the need to get the model to do something specific, such as predict (with high accuracy) behavioural or neural data. Templates are useful for this kind of work because they provide tractable and usually well-understood structures from which to build such models. She states: “What makes [templates] popular is their tractability and solvability, which reflects the results-orientedness of modelling: the starting point is often the output and effects that models are supposed to produce” (Knuuttila, 2011, p. 268). As noted above, this contrasts strongly with the description of model-based science provided by Weisberg and Godfrey-Smith, which is more exploratory in nature.

However, at least in cognitive neuroscience, models built from templates must also represent their targets in a non-trivial way. Models must be ‘plausible’, or sensitive to ‘bottom-up’ and ‘top-down’ constraints. Bottom-up constraints are those from the level of implementation, such as research from molecular and systems neuroscience. Research on what goes on at the neural level is still in fairly early days, but it determines boundaries around what the system does or is capable of, and how abstract processes can be implemented. Top-down constraints come from assessments of what the system is likely to do, in ecological, temporal, thermodynamic, evolutionary or other)terms. For example, if a model of a cognitive process is computationally intractable, and insufficient means are given to transform it into a tractable one, then the model is essentially implausible (van Rooij and Wareham 2012)[3].

Together with plausibility, the idea of a correction set is strongly tied to Weisberg and Godfrey-Smith's notion of construal noted earlier. Researchers must identify properties of the target system that they think should be included in a model (scope) and choose a template accordingly. In translating a template, researchers must identify those parts that they need to be translated and those that can be left out, or will play only a minor role in the model (assignment). Since models are constructed to accurately predict the behaviour of the system, and to satisfy top-down and bottom-up plausibility constraints, they will also satisfy the fidelity criteria to a greater or lesser degree.

The way these features are exhibited in modeling practice in cognitive neuroscience are described briefly below, with examples from research on decision making, followed by a discussion of how these practices constitute a different form of model-based theorising to that offered by Weisberg and Godfrey-Smith.

Step 1: Choice of target and template

Cognitive neuroscientists interested in decision making now use experimental paradigms from psychophysics, psychology, neurophysiology and economics to shed light on how decision making is carried out at the neural or systems level. For example, visual displays of moving dots (random-dot motion), for which subjects have to 'choose' which of two directions most of the dots are moving, are now also used to investigate sensory decision making because they are thought to exhibit similar computational principles to 'higher level' decision making. These paradigms are ideal targets for modeling because they are often quite simple, incorporate a narrow range of variables, and are easy to manipulate and control. Whether the paradigms used as the target phenomena for these models are in fact 'representative' of decision-making in general is an open question, but one that researchers are optimistic about:

"Our scope is somewhat narrow: We consider primarily studies that relate behavior on simple sensory-motor tasks to activity measured in the brain because of the ability to precisely control sensory input, quantify motor output, and target relevant brain regions for measurement and analysis. Nevertheless, our intent is broad: We hope to identify principles that seem likely to contribute to the kinds of flexible and nuanced decisions that are a hallmark of higher cognition." (Gold and Shadlen, 2007, p. 536)

Modelers must also choose which template they want to use to build their model. Modelers often give theoretical justification for using particular templates, by arguing that they have the right sort of structure, or the right sort of variables, given the

[3] This often comes out in discussions within the discipline of the need for cross-level constraints (e.g. Forstman et al. 2011, Jones and Love 2011), and relatedly on whether paying particular attention to top-down or bottom-up constraints is more likely to lead to models that capture general principles of cognitive processing (e.g. Griffiths et al. 2010, McClelland et al. 2010).

bottom-up or top-down constraints on the target system. For example, in studying sensory decision making, Gold and Shadlen (2007) justify their use of a Sequential Analysis framework by arguing that it satisfies a top-down constraint on the way that brain mechanisms are likely to work, given temporal features of cognitive processing:

"…SA [sequential analysis] includes a termination rule. Analogous mechanisms in the brain are required to make decisions and commit to alternatives on a time frame that is not governed by the immediacy of sensory input or motor output, a hallmark of cognition." (p. 542)

Other modelers also use neural evidence to satisfy bottom-up constraints, and to emphasise the likelihood that particular key variables in the model really are represented in the brain. For example, Corrado et al. (2009) promote the use of reinforcement learning models (originally from machine learning) in decision making research, which provide normative models of how agents should act in a given uncertain environment in order to maximize reward. A key variable in these models as used in decision making is reward prediction error, which is the difference between the expected and actual reward. As support for using this kind of model, Carrado et al. (2009) state that:

"The correspondence between the error signals [reward prediction error] postulated by RL [reinforcement learning] models and the phasic responses of midbrain dopamine neurons recorded in simple conditioning tasks has led to the proposal that RL might reflect the literal mechanism through which the brain learns the values of states and actions." (p. 477)

The need for such justifications can be seen to come partly from the use of existing experimental paradigms as target phenomena for model-based approaches. These target phenomena often come already associated with a range of theoretical concepts and principles from existing behavioural models, and indeed particular computational models for predicting behavioural outcomes. Yet there is often doubt that these principles and models are appropriate for finding out about the *neural* processes by which decision making is implemented. For example, while economic models based on some variant of expected utility theory may adequately capture group behavioural data, there is space to doubt that they can describe the neural underpinnings of these behaviours (see e.g. special issues on neuroeconomics in *Economics and Philosophy* 24:3, 2008, *Journal of Economic Methodology* 17:2, 2010). In this case, the choice of a computational template often requires far more justification than in other cases of model-based science in order for the resulting model to be seen as fit for its purpose.

Step 2: Model construction

As described in more detail above (and later in Section 5), correction sets are then generated to translate a computational template into a model, based on a range of background knowledge about the template and the target. Importantly, this stage involves the elucidation of the ways in which the resulting model 'maps' to a target. Parts of the target may be left out, and parts of the template may not be fully interpreted or seen as particularly important, depending on the theoretical claims that the model is being used to make and test.

Step 3: Model analysis and testing

In contrast to the cases of model-based theorising described by Weisberg and Godfrey-Smith, general principles are not typically the sort of things that come out of analysing models in cognitive neuroscience. This largely stems from the use of computational templates. Templates provide uninterpreted principles, and templates are chosen precisely because modelers think that these principles can capture something about cognitive processing, be this Bayesian inference, sequential sampling, parallel distributed processing, and so on. In this case, general principles are not novel products of template-based modelling in cognitive neuroscience, as the identification of appropriate general principles is precisely what drives the choice of particular template.

Instead, a common aim in model-based cognitive neuroscience is to use a model to investigate whether, and to what extent, these general principles really do apply to cognitive processes, to better understand the scope of these principles, and how they might be implemented. Modelers constructing reinforcement learning models aim to see how well this kind of model can correctly predict behavioural data from a particular experimental paradigm, and to test how well the values of reward prediction error in the model correlate with neural activity in some area of interest. From here it is possible to make inferences about whether or not (and where) reward prediction error, or something like it, is actually represented in the brain, perhaps what sort of reinforcement learning algorithm is computed in the brain, and potentially about the architecture of relevant brain systems. The generation of this kind of theoretical knowledge is very different to the analysis of formal properties of an uninterpreted abstract framework or template.

Models are also tested in multiple ways, at multiple steps of the process of model-based theorising. Following the justifications built into the construction of a model (e.g. justifying the choice of target, template, correction set), the fit between the model and real data is a crucial way to test model-target fit. Corrado and Doya (2007, p. 8180) suggest two ways in which this is done: predictive and generative tests. The predictive accuracy of a model describes how well the model can predict a subjects' next choice given their decision history. This is similar to a cross-validation test, but on a smaller scale. The model is fitted to a sequence of choices, and must be able to predict (to some degree of accuracy) the next choice. In contrast, generative tests show how well a model can generate realistic choice patterns over long time periods. This is essentially to use the model as a simulation.

It is also possible to use neural or imaging data to distinguish between competing models. When both models use the same basic idea (e.g. reinforcement learning), but where the variables in the model take very different values, neural activity known to be related to reinforcement learning processes can be used to tell which model better captures the underlying processes (see Hampton et al., 2006, for possibly the first use of this in neuroeconomics).

A further factor relating to model testing is parameterisation and comparative testing. Parameters are free variables whose values determine the behaviour of a system, and they can be set by modelers to maximise the predictive accuracy of the model. However, many models, perhaps particularly those based on computational templates, contain fictionalised variables and relations that can reflect part of a template that is unlikely to be actually realised in the particular system under study, or just consist of 'fudge factors' put into the model to get it to work properly. The problem comes when a large number of fictionalised parameters are present in a model, as the predictive accuracy of the model may be due to these parameters, rather than the variables that modelers think actually map to some feature of the target system. To try to address this problem model comparisons now often include controls for model complexity (i.e. controlling for numbers of free parameters, see e.g. Pitt and Myung 2002; Roberts and Pashler 2000).

And around again

Steps 2 and 3 are used to iteratively refine models, to update correction sets and identify the scope of application of the model. Considerations of the 'fit' between the model and the target are present in all steps of modelling, from the choice of target and template, the translation of the template using a correction set, model construction, development and testing. Knowledge of a model's fit to target systems is used to justify and delimit the theoretical claims made by modelers in cognitive neuroscience, and it is on these grounds that models are evaluated by others.

4: What Kind of Representation?

The question then becomes whether this kind of model-based theorising really involves direct or indirect representation. I suggest that the answer is on the side of direct representation, such that appropriate analysis of models is analysis of the target, when modellers know that their model appropriately represents a target. This is essentially no different to the practices involved in ADR. That is, although model-based theorising in cognitive neuroscience and ADR go about ensuring model-target fit in different ways, they are arguably following the same goal of constructing a theoretical structure that represents, in relevant respects, for relevant purposes, a target. In this case, when models represent in model-based cognitive neuroscience, they do so directly.

The basis for this claim stems from the practices surrounding the use of computational templates in cognitive neuroscience discussed above. The kind of abstract model analysis described by Weisberg and Godfrey-Smith, where general patterns of behaviour are uncovered, and then compared with real phenomena, is just not part of this modelling practice. Instead, the analysis and identification of general principles is already done by those who have developed computational templates, which are then utilised by other modelers. Templates are then chosen and interpreted in particular ways because there is already some reason to think that the general principles found in the template can be used to represent core features of cognitive processing.

From this, Weisberg's (2013) note of the "less formal determination of representational capacity" (p. 75) present in the early stages of modelling, sets up some basic representational relations between the general principles in the model and the target system. Theoretical claims based on these initial representational relations may very simple, and modellers will of course be less sure about whether (and how) other parts of the model represent features of the target. But over time, modelers get a better understanding of what they can use the model to represent, and what not, and so what kind of theoretical claims they can make. The key here is to see the representational capacity of the model as changing over time, over periods of testing and development. The representational capacity does not change from indirect to direct representation, but as gaining representational capacity from the first (generic) claims, to more and more detailed claims about how the model represents a target in relevant respects, for relevant purposes.

This might make it sound like theorising in cognitive neuroscience is ‘simply’ a matter of fitting a template to a target system, making it equivalent to Step 3 of Weisberg’s account of model-based theorising, and therefore merely a case of indirect representation after all. The story is little more complex however, and points to the need to contrast some apparent differences between the development of theoretical structures under the ADR framework, and under the modelling approach described here, in order to better understand how researchers treat the representational capacity of their theories.

First, in ADR, (according to the examples in Weisberg and Godfrey-Smith), theorising may more often be shallow, in the sense that theoretical structures describe patterns in data rather than causal mechanisms that generate those patterns. So, Mendeleev’s periodic table organised elements according to abstract properties that were already known about, but could not offer deep explanations about the behaviour of chemical elements. In contrast, modeling is often used to suggest possible or plausible causal mechanisms that explain the existence of these patterns. In cognitive neuroscience in particular (as discussed further below), models are used to bridge the gap between stimuli (input) and behaviour (output), and specify possible cognitive computations or processes. So, one apparent difference between ADR and modeling might be in terms of the ‘depth’ of theorising, with ADR being more aligned with surface patterns in data, and modeling more aligned to making novel claims about underlying, and as yet unobserved, structures or causal mechanisms.

Second, theorising in ADR may also rely less on existing theoretical knowledge, and somehow take its inspiration more straightforwardly from data. In modelling work in cognitive neuroscience, theoretical knowledge about templates, and background empirical and theoretical knowledge about cognitive systems, is used to construct a model. It might therefore be the case that ADR is driven more by data than by theory, and so again represent its targets more directly.

Third, models (particularly early on in their development) can display the peculiar feature that they can often include terms for which there is no clear interpretation, which are obviously fictional or as yet not fully analysed or specified. This may be less true of theoretical structures developed under ADR that stick firmly to the data, and so again more directly represent their targets.

There are several senses of ‘directness’ present here, not all of which always side with ADR. First, theorising according to ADR does not always shallowly track patterns in data, but ideally goes beyond what is present to make interesting and novel claims about what is not ‘directly’ observed. Models may do this more obviously, but it is not a practice only found in modeling. Second, theorising according to ADR can often draw on a range of theoretical knowledge in the same way as modeling. Darwin’s theories about atoll formation, and Maynard Smith and Szathmáry’s claims about evolutionary transitions drew extensively on geological and biological theory. Third, ADR can produce theories with untested, vague or fictional variables just as modeling can, particularly if they are in the business of theorizing about things not immediately exhibited in the data. Finally, implicit in some of the discussion of ADR and modelling is the idea that modeling is often more abstract (particularly true of templates). Yet it is not clear in what sense a highly abstract theoretical structure less 'directly' represents its target, particularly in the sense that Weisberg and Godfrey-Smith are driving at. To the extent that all theorizing involves abstraction, representational relations are always some distance away from (models of) data.

What is in common with some of these (misleading) distinctions, and what indeed seems to drive intuitions about the 'directness' of representation in ADR, as compared with 'indirectness' with modeling, is that theoretical structures in ADR are often presented as complete and 'correct'. In contrast, models are presented as incomplete and in need of further testing and development. Yet across both ADR and modelling in cognitive neuroscience, no theorisers treat their theoretical structures as successfully and wholly representational from the start, but as better or worse representations of (parts of) targets, for different purposes, that require vigorous testing. The basic development of knowledge of if, and how, theoretical structures represent their targets is the same: it grows over time, and stems from testing and theory development. So, initially, those engaged in both ADR and the kind of modeling described above may formalise and analyse their representations independently of the target phenomenon, in order to make predictions and specify the scope of the theory. This in done in the hope, but perhaps not the belief, that such analysis can tell us about the target system. Further, both models and other theoretical structures can contain abstractions, idealisations and fictions, and researchers of all stripes need to specify how their theories apply to target systems, given these 'false' components.

In this case, both are engaged in the same job of mapping out, testing and developing the representational capacities of theoretical structures, where representation is partial, graded and purpose-relative. And for both ADR and the kind of model-based science discussed above, any successful representation is direct representation. Models and other theoretical structures in cognitive neuroscience are constructed and developed in order to give knowledge about target systems, not in order to give knowledge about formal features of the theory itself.

All this is to say that using a pre-existing computational template to theorise about a target system can be just as involved a theoretical process as engaging with ADR. While the approach in model-based neuroscience is obviously different from that of ADR, it is not so different in fundamentals; models here are still based on abstractions from specific target systems and sets of data, though highly scaffolded using existing computational or mathematical structures. Models are meant to represent targets (in relevant parts, in relevant ways), and analysis of appropriate, representational parts of models is treated as analysis of the target. Discussing only early stages of modeling and later stages of ADR obscures the fact that both kinds of theorizing include long stages of getting things wrong, testing, and changing and refining theoretical structures in order to learn exactly what their direct, partial and purpose-relative representational capacities are.

5: Theorising and Robustness

An example is given below that further illustrates the nature of partial, abstract and direct representation often found in modeling work, that draws on ideas expressed by modelers themselves. This also sheds some light on the role of robustness in this kind of model-based theorizing, which warrants some otherwise peculiar theoretical inference structures.

Corrado and Doya (2007) use the template of reinforcement learning to make theoretical progress on the neural basis of decision making, but only treat certain parts of their models as (currently) representational of parts of their target system. They outline a ‘model-in-the-middle’ method that breaks the tradition of directly correlating neural data with stimuli and behavioural patterns (ADR). Instead, models are used to theorise about the kind of computations that the brain performs in a task, and neural data is correlated with the intermediate, internal variables described in the model. This is a clear case of using modeling to scaffold theoretical work for which ADR is simply not amenable, at least with the current state of cognitive neuroscience. They state that this method is particularly useful because models make it possible to calculate exact values of variables that are not immediately obvious from neural data, and to uncover effects that would otherwise go unnoticed from mere inspection of neural data (see Corrado and Doya pp. 8179-8180).

However, the treatment of the representational capacities of these models can initially be puzzling for the purposes of finding out anything about a target system:

“The essential function of the model here is not necessarily to serve as an explicit hypothesis for how the brain makes a decision, but only to formalize an intermediate decision variable. Thus, even if a model that contains some components that seem unlikely to be literally implemented in the brain…the resulting proxy variables may prove useful if they correctly capture the essential concept that the researcher wishes to study…” (Corrado & Doya, 2007, p. 8180)

Here, the authors suggest various things about how they treat the representational capacity of their model. First, they do not think that the computation described in the model actually describes a computation going on in the brain, nor that the computation is itself the focus of using the model (they do not take it to be a hypothesis). Second, the model contains components that are unlikely to be implemented in the brain (perhaps idealisations or fictions). So, the authors seem to be using a model that contains components and a computation whose representational accuracy they are almost entirely agnostic about, to make predictions about choice behaviour and neural activity in fairly detailed form, over time, and for individual subjects, for a single variable in the model (reward prediction error). It is not clear if this interpretation of the reinforcement learning template is warranted as a way of making theoretical progress about the neural basis of decision making. One possible response to this is offered by Mars, Shea et al. (2012):

“…[this strategy] perhaps reflect[s] the fruitful neutrality that modeling permits about the relation between model and target. That allows Corrado et al. to remain neutral about how, or even whether, various aspects of a model will map onto computations performed in neural circuits, while still discovering how other aspects of the model (e.g., prediction errors) map onto the target system.” (p. 259)

However, this response seems inadequate, as it does little more than restate the problem. Clearly there are cases where modelers can reserve judgement about some details or components of a model and still learn about a target system, but at least in the case above, there seems little reason to think that the assessment of the representational capacity of the model described by Corrado and Doya warrants confidence in the inferences they make from it. What are needed are some reasons to think that the model is representationally appropriate for the kind of inferences that are drawn from it.

One solution is to see the agnosticism about the model above as a kind of local agnosticism. So, one might accept that the model can represent a target in so far as the target engages in some kind of reinforcement learning, even if detailed parts of the model are not treated as representing parts of the target. As Humphreys (2002) notes:

“…the justification of scientific models on theoretical grounds rarely legitimates anything more than such a parametric family, and the adjustments in the fine structure by empirical data are simply an enhancement of the empirical input traditionally used to pin down the models more precisely.” (p. 9)

Latent in this treatment of the representational capacity of models is a notion of robustness (Levins 1966, Wimsatt 1981, Weisberg 2006, Weisberg and Reisman 2008). Robustness is (roughly) invariance over something; over input, or assumptions, or measurements, and so on (see Woodward 2006 for discussion). In contrast to recent discussions on the role of robustness in confirming or lending confidence to models (Kuorikoski et al. 2010, 2012; Odenbaugh and Alexandrova 2011), the claim here is that robustness can enter much earlier into the process of model construction. This is because knowledge of the robust properties of a template plays a large role in template selection. It is typically the robust properties of a template that constitute the general principles of the template, and these are precisely what guide the initial selection of a template. In addition, an important way of evaluating cognitive models is how well they track cognitive performance when subjects perform tasks well, even under pressure, but also when they fail, sometimes disasterously. In this case, cognitive models need to be robust and fragile in the same ways in which cognitive processing is robust and fragile.[4]

So, in choosing between templates, modelers rely on their knowledge of the range of input or boundary conditions for which a template will generate a particular pattern of output, and when it will start to degrade (termed parameter robustness/sensitivity analysis). They know the variables that can be ignored or altered without altering the main patterns of outputs, and which ones cannot (structural robustness). They also know what kinds of formal tools they can use to express the model that preserve the patterns of output (representational robustness, see Weisberg 2006 for more on these

[4] For example, connectionist models provided better a better theoretical structure than symbolic models for explaining errors made during language learning, and in degradation of cognitive performance due to brain lesions (e.g. Thomas and McClelland 2008).

distinctions). This means that modelers can use their knowledge of a template's robust properties, and knowledge of a target system's robust properties (as evidenced at the behavioural level), as warrant to try out different versions of a model that preserve the robust properties of the original template.

In this case, the particular computation found in the model need not serve as a hypothesis about the specific computation going on in the brain, and modelers need not commit to the representational capacity of all or most variables in their model. However, in doing this they are safe in the knowledge that at a very abstract level, the model represents a core abstract feature of the target; that it works according to some sort of reinforcement learning algorithm. They can use this knowledge to make appropriate theoretical inferences about the target, given the robust properties of both, and the representational relations that hold between them. Over time, given continual model testing and model development, knowledge about the representational relations between model and target warrant other theoretical claims.

This further illustrates that throughout the kind of model-based theorising found in cognitive neuroscience, modelers treat representational relations between their model and target systems as partial, purpose-relative, and in need of testing. Theorising under ADR works in essentially the same way. Inferences made in both types of theoretical structures are assumed to hold in the target in virtue of the theory directly representing the target in relevant parts and in relevant ways. Both cases therefore exhibit direct representational relations.

6: Conclusion

The examples above show that model-based theorising is not always of the variety described by Weisberg and Godfrey-Smith. Both Godfrey-Smith and Weisberg recognise that there may be a plurality of methods of theorising, and Godfrey-Smith (2006) in particular notes that modelers may shift between direct and indirect uses of models (pp. 734). However, the sections above were aimed at showing that in some types of modeling, as distinct from the generation and analysis of uninterpreted computational templates, there is a way of reading modeling practices as involving direct representation, if only in a very partial kind and at an abstract level. In particular, this seems to be a good way of characterising the progression of model-based theorising in cognitive neuroscience.

Much of this re-evaluation depends on treating ADR as just as messy, dynamic and error-prone a process of theorising as modeling is. Theories developed in ADR are not guaranteed to represent their targets appropriately, and so analysis of these theories does not always translate successfully into analysis of targets. Theories of any kind require ongoing testing and development. While the structures found in model-based theorising may initially be borrowed from elsewhere, the intentions of researchers to provide representations of target systems, and the ways of identifying a theory's representational capacity are the same.

The use, rather than construction, of computational templates drives many of the differences explored above. In this case, the kind of model-based theorising described here is likely to be found in other areas of research that also use templates to make theoretical claims, which according to Humphreys (2004) and Winsberg (2010) are ever growing. At the very least, this suggests that further investigation is needed into the uses of computational templates in scientific research and their roles in theory building.

As a final note, the arguments made above could potentially also be used to claim that Volterra and Maynard Smith and Szathmáry's method of theorising involves direct, rather than indirect representation. Analysis of their models is supposed to tell us about predator-prey systems and evolutionary transitions without a huge amount of comparison between these models and detailed biological data. This is presumably because the initial (simplistic, generic, abstract) representational relations between the model and classes of target phenomena are deemed sufficient for treating some analyses of the models as analyses about (generic, abstract) features of a class of targets. In one sense, this could be treated as direct representation of a very abstract kind. This stems from the core idea sketched here, that questions uniquely asked about models (representation, whether explanations can false parts) can also be asked about theories that stem from traditional practices such as ADR, and this may in some ways make them more similar than different.

Acknowledgements

Many thanks to Matteo Colombo for ongoing (friendly) criticism, Catherine Stinson for discussion on model-related matters, the audience at the 'Explanatory Power II: Understanding through Modelling?' workshop in Bochum where part of this paper was presented, to the Philosophy of Neuroscience group in Tuebingen, and to two anonymous reviewers for their extremely helpful comments.

References

Bayer, H.M. and Glimcher, P.W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47, 129-141.

Corrado, G. and Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *The Journal of Neuroscience*, 27, 8178-8180.

Corrado, G., Sugrue, L., Brown, J. R., and Newsome, W. (2009). The trouble with choice: Studying decision variables in the brain. In (Eds. P. Glimcher, Camerer, C. F., Fehr, R. and Poldrack, R. A) *Neuroeconomics: Decision Making and the Brain*, pp. 463-480. Academic Press.

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., and Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and forma cognitive models: Opposites attract? *Trends in Cognitive Sciences*, 15, 272-279.

Friston, Karl. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1456 (2005): 815-836.

Giere, R. N. (1988). *Explaining Science: A Cognitive Approach*, Chicago: University of Chicago Press.

Glimcher, P. W. (2011) Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Science of the* USA, 108, S3, 15647-15654).

Glimcher, P. W., Dorris, M. C. and Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and Behavior*, 52, 213-256.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, 21, 725-740.

Godfrey-Smith, P. (2009). Model and fictions in science. *Philosophical* Studies, 143, 101-116.

Goodwin B. C. (1963). *Temporal Organization in Cells.* New York: Springer.

Gold, J. I. and Shadlen, M, N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535-574.

Griffiths, T. L. Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357-364.

Hampton, A.N., Bossaerts, P. and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26, 8360-8367.

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Humphreys, P. (2002). Computational models. *Philosophy of Science*, 69, 1-11.

Jones, M. and Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Brain and Behavioral Sciences*, 34, 169-188.

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science*, 42, 262-271.

Knuuttila, T. and Loettgers, A. (2011). The productive tension: Mechanisms vs. templates in modeling the phenomena. In (Eds. P. Humphreys and Imbert, C.) *Models, Simulations, and Representations*, pp. 3-24. Routledge.

Kuorikoski, J., Lehtinen, A. and Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science*, 61, 541-567.

Kuorikoski, J., Lehtinen, A. and Marchionni, C. (2010). Robustness analysis disclaimer: Please read the manual before use! *Biology and Philosophy* 27, 891-902.

Loettgers, A. (2007). Model organisms, mathematical models, and synthetic models in exploring gene regulatory mechanisms. *Biological Theory*, 2, 134-142.

Levins, R. (1966). The strategy of model-building in population biology. *American Scientist*, 54, 421–431.

Lotka, A. J. (1920). Undamped oscillations derived from the law of mass action. *Journal of the American Chemical Society,* 42 1595-1598.

Mars, R. B., Shea, N. J., Kolling, N. and Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *The Quarterly Journal of Experimental Psychology,* 65, 252-267.

McClelland, J. L., Botvinick, M., Noelle, D. C., Plaut, D. C., Rogers, T.T., Seidenberg, M. S., and Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences, 14,* 348-356.

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. *The Journal of Neuroscience*, 32, 551-562.

Odenbaugh, J. and Alexandrova, A. (2011). Buyer beware: robustness analyses in economics and biology. *Biology & Philosophy,* 26, 757-771.

Ortega, P. A. and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Academy A*, 469, 20120683, doi: 10.1098/rspa.2012.0683

Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science*, 6, 421-425.

Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358-367.

Roughgarden, J. (1979). *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. New York: Macmillan Publishing Co.

Sanfey, A. G., Loewenstein, G. McClure, S. M. and Cohen, J. D. (2006). Neuroeconomics: cross-currents in research on decision-making. *Trends in cognitive sciences,* 10, 108-116.

Scholl, R., and Räz, T. (2013). Modeling causal structures. *European Journal for the Philosophy of Science*, 3, 115-132.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.

Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology*, 7, 1-31.

Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, 70, 3-27.

Thomas, M. S. C. & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed). *Cambridge handbook of computational psychology*. Cambridge University Press. 23-58.

van Rooij, I. & Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *Journal of Mathematical Psychology, 56*,232-247.

Weisberg, M. (2006). Robustness analysis. Philosophy of Science, 73, 730-742.

Weisberg, M. (2007). Who is a modeller? *British Journal for the Philosophy of Science*, 58, 207-233.

Weisberg, M. (2010). Target direct modeling. *The Modern Schoolman*, 87, 251-266.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Weisberg, M. and Reisman, K. (2008). The robust Volterra Principle. Philosophy of Science, 75, 106-131.

Wimsatt, W. C. (1981). Re-Engineering Philosophy of Limited Beings: Piecewise Approximations to Reality. Harvard University Press.

Winsberg, E. B. (2010). *Science in the Age of Computer Simulation*. Chicago University Press.