

# Model-Based Science and Artificial Cognitive Systems: The Philosophy of Computational Modelling

Terrence C. Stewart <terry@ccmlab.ca>

Carleton Cognitive Modelling Lab, Institute of Cognitive Science, Carleton University

## **Abstract**

This paper applies *model-based science* to the domain of computational modelling of cognitive systems. The central claim is that we can treat computational models in the same manner as traditional science has treated mathematical models. This has a direct implication in terms of how we should interpret the results of investigations involving artificial cognitive systems – the models we create can be shown to be both predictive and (eventually) strongly equivalent to real biological systems. However, in order to reach such conclusions, we must address the increased complexity of the computational models. A methodology for doing so is presented.

## **Model-Based Science**

In *Computation and Cognition*, Pylyshyn (1984) argues that cognitive models should “strongly equivalent” to the real systems being modelled. That is, they should not only give the same outputs for the same inputs (i.e. exhibit the same behaviour), but they should also use the same algorithms within the same functional architectures as the real systems. However, as he describes, this is a difficult and long-term task, relying on converging evidence from a variety of sources. This should include information such as reaction times, cognitive penetrability, and lesion effects; all aspects of the cognitive system which are not traditionally considered part of its behaviour.

This requirement of having converging sources of evidence is, unsurprisingly, similar to that seen in the physical sciences. Indeed, it is exactly when a theory accounts for a wide variety of data in disparate domains that physicists start to refer to it as a Law. Importantly, newer theories which account for even wider ranges of phenomena can succeed older theories. This does not, of course, mean that the older theories were not useful. On the contrary, these older theories give a concise representation of *what is currently known*, and are a necessary part of the development of the science. We argue herein that the creation of artificial cognitive systems can proceed in a similar manner: from simple models of a limited set of behaviours in limited situations to more complete models.

Our core thesis is that the scientific framework used by physicists to create mathematical models is directly applicable to researchers creating computational models of cognitive behaviour. This forms the basis of how we ought to use computational models within scientific investigation. However, due to the increased complexity of such models, certain aspects of the detailed methodology must be adapted. The result is an explicit methodology which addresses many of the current concerns surrounding the use of such models. This includes the issues of parameter fitting, model comparisons, and the epistemic status of the entities postulated by such models.

## **Modelling in Physics**

If we are to interpret computational modelling as being the same as mathematical modelling, we need to first closely examine the use of mathematical models. The goal of this extended digression is to form a detailed understanding of how this model-based way of thinking about science applies to a straight-forward and well-understood domain such as basic physics. Once this depiction is made clear,

we will then apply it to the more complex situation of computational models within cognitive science. The fundamental claim is that the scientific methodology will not change between the two situations.

For our example, we consider Newton's theory of gravitational attraction, summarized by

$$F = \frac{G \cdot m_1 \cdot m_2}{d^2}$$

This formula allows us to determine the force of attraction ( $F$ ) between two objects of known mass ( $m_1$  and  $m_2$ ), separated by a particular distance ( $d$ ). To make use of this theory, we use the formula to create a predictive mathematical model, *customized to the particulars of the situation to which we are applying the theory*. Since this is a mathematical model, we do this by replacing the variables within the model with the relevant values and performing the indicated calculation.

For example, if we have a 5 kg object near the surface of the Earth, then we can let  $m_1$  be 5 kg,  $m_2$  be  $5.9742 \times 10^{24}$  kg (the mass of the Earth),  $d$  be  $6.3781 \times 10^6$  m (the distance from the surface of the Earth to its center), and  $G$  be  $6.6730 \times 10^{-11}$  m<sup>3</sup>/kg s<sup>2</sup> (the universal gravitational constant). The result is 48.999 kg m/s<sup>2</sup>, which is the model's prediction of the force downwards on the object due to gravity.

Testing this prediction is difficult, since directly observing *force* is not possible. However, Newton's second law of motion defines the relationship between force and acceleration ( $a$ ).

$$F = m \cdot a$$

If the Earth's gravity is the only significant force acting on the object, then the acceleration can be predicted as well. Mathematical rules allow us (in certain situations) to reorganize a mathematical model into other forms, while ensuring that the predictions remain identical. Thus, we can change the model to the following form.

$$a = \frac{F}{m}$$

We again substitute the appropriate values into the equations, resulting in a predicted acceleration of 9.8 m/s<sup>2</sup>, which is in accord with the observed behavior of the real object.

While this description of the process of using mathematical models is relatively typical, extending this example to computational modelling requires further examination of certain aspects.

### **Parameters (Input)**

To begin, more must be said on the determination of the particular values being used for substitution. How do we know that the mass of the Earth is  $5.9742 \times 10^{24}$  kg, or that its radius is  $6.3781 \times 10^6$  m? Indeed, what does it mean to say that the object has a mass of 5 kg? Interestingly, using only geometry and the position of the sun at the same time in two different cities, the ancient Greek scholar Eratosthenes was able to determine the Earth's radius. The mass of reasonably sized objects is determined by defining one particular reference object to be 1 kg, and then using a balance to determine that the mass of a given object is, in this case, five times the mass of the reference object, or 5 kg. This method, however, is not applicable to determining the mass of the Earth.

Indeed, for 120 years after the development of the theory of gravitation, the mass of the Earth was unknown. Importantly, *the theory was useful without knowing absolute values of all the parameters*. Due to the simplicity of the relationship between  $G$ ,  $m_1$ , and  $m_2$ , they can be combined into a single parameter, and for any given situation, we can determine what value for this combined parameter best 'fits' the particulars of a given situation. For example, if we use the formula to predict the influence of the Sun on Jupiter, we do not need to know the mass of either. Instead, we can observe the path of Jupiter for a short period of time, determine the force applied by the Sun that would be required to result in such a path<sup>1</sup>, and then determine for what value of this combined parameter the model would give the same result. Once this value is determined, it can then be used in all future predictions of the gravitational influence of the Sun and Jupiter. In other words, *the model parameters are fit to one situation, and then applied to other situations*.

However, if the mass of the Earth was known, then we would be able to apply the formula without this initial fitting stage. This can, of course, be seen as doing parameter-fitting on the individual parameters in the formula, instead of the combined context-dependent parameter. Knowing these values would allow for a more rigorous test of the theory, and would lead to more general uses. In particular, it would give us a value for  $G$ , the gravitational constant. This is a parameter which has *the same value in every application*. In fact, in terms of using the theory to produce models, we can consider that the theory itself *includes* the specification of this parameter. Taking this approach, James Cavendish evaluated  $G$  by devising a situation where the gravitational attraction between two reasonably sized objects could be measured<sup>2</sup>, and used this result to determine the best-fitting value of  $G$ . This, in turn, could be used to determine a suitable value for the mass of the Earth. Importantly, we must note that these evaluations are done *assuming the theory is correct*. They give us values that can be used in the creation of other models in other situations. These other models can then be considered predictions of the underlying theory, which includes the specifications of these constants.

### **Processing (Calculation)**

A further important consideration is masked by the idea of predicting the *acceleration* of an object due to gravity. Much like force, *acceleration is also not directly observable*. However, the general framework in which the theory of gravitation is embedded defines acceleration as the change in velocity over time, and defines velocity as the change in position over time. Thus, given a known acceleration, we can predict the position of an object over a period of time, and given known positions, we can determine what pattern of acceleration would give the same result.

For simple cases, where the acceleration does not change significantly, the positions can be determined by basic mathematical manipulation of the definitions of the terms. If  $v_0$  is the initial speed of an object, and it accelerates at  $a$ , then the resulting velocity ( $v$ ) and distance traveled ( $d$ ) at a particular time  $t$  are given by the following formulas.

$$v = v_0 + a \cdot t \qquad d = v_0 \cdot t + \frac{a \cdot t^2}{2}$$

However, if the acceleration due to gravity *changes* significantly over time (as it would for planets or comets orbiting the Sun), then the matter is more complicated. The issue is that the acceleration is

---

1 Using geometry and Newton's laws of motion.

2 Using a large barbell suspended by a string placed near other weights.

due to the force of gravity, which is dependent on the distance between the objects. The acceleration causes *changes to the distance*, which in turn affects the acceleration.

This feedback loop leaves matters in a strange situation. We have a mathematical model of the phenomenon, *but we may not be able to use mathematics to determine its predictions*. The theory of gravitation allows us to create models of the acceleration of Jupiter due to the Sun at different points in time, but how can these be combined to produce an overall path for Jupiter?

The approach for dealing with this problem is to divide the overall period of time of interest into small intervals. Within each of these time intervals, we assume that the acceleration is constant. We can assume that the force due to gravity will not change significantly within that amount of time, and so we can make use of the simple formulas to determine the position and velocity during that interval. Then, given the new distance, we can use the gravitational formula to determine the new acceleration for the next interval.

Performing the calculations in this manner would clearly require inhuman patience and care. Furthermore, the results would change depending on how large the intervals were chosen to be. Instead, the theory of gravitation was accompanied by the mathematical technique known as *calculus*. The idea here involves directly determining what the answer would be *if the intervals were reduced to an infinitely small size*. This remarkable short-cut then solves our problem, allowing the predictions to be determined without a tedious process of gradual calculation.

Suppose that at a given time the position and velocity of a planet can be determined, and that the force is known. Then, according to Newton's laws we know the change in velocity during a short time interval. Knowing the initial velocity and its change, we can find the velocity and position of the planet at the end of the time interval. By a continued repetition of this process the whole path of motion may be traced without recourse to observational data. This is, in principle, the way mechanics predicts the course of a body in motion, but the method used here is hardly practical. *In practice such a step-by-step procedure would be extremely tedious as well as inaccurate*. Fortunately, it is quite unnecessary; mathematics furnishes a short cut, and makes possible precise description of the motion in much less ink than we use for a single sentence. (Einstein and Infeld, 1938, page 31, emphasis added)

However, the calculus technique *only works for certain situations*. If we consider a model of three planets interacting due to gravity, then the mathematics are intractable, and there is no known short-cut. Fortunately for physicists, there tends to be only one dominant gravitational force for most celestial bodies (the force of gravity between the Earth and Jupiter is much less than between the Sun and Jupiter, so the effects of Earth and the other planets can safely be left out of such a model).

While there are a number of attempts at mathematical tricks for dealing with special cases of this 'three-body problem', in general, the only solution is to follow the original approach of dividing the time into small components and performing iterative calculation. While the result of this is dependent on the size of the time interval, there are methods for determining how large such a variation could be, and so the amount of uncertainty due to this approach can be known. This is fortunate, since *most complex models are mathematically intractable*. As models get more complicated, this iterative technique is required for determining their temporal predictions.

## **Matching (Output)**

The final aspect of mathematical modelling which bears closer examination is the process of confirming the model's predictions. If a model predicts that an object will fall 4.9 meters in one second, and measurement reveals that it only fell 3 meters, then it may be reasonable to conclude that the model was not appropriate for the situation. If, however, the object is observed to fall 4.899999 meters, then it seems our model is supported. Between these extremes, it seems that a principled method is needed for determining if the model matches or fails to match the observed situation.

In some situations (especially when the iterative approach is used for determining the predictions over time), there is a method for mathematically calculating the maximum 'error' possible due to choices such as the size of the time interval. This can give a lower bound on the allowable difference between the prediction and the final result. Similar considerations can be made due to the variability in the initial measurements used to set the parameters of the model (such as the object's mass).

However, in practice, there is another source of 'error' between model and reality: the influence of factors outside the scope of the theory. The simple model of gravitational force on an object only gives accurate results *if there are no other significantly large forces*. When there are other such considerations, we must build more complex models involving other physical laws as well (such as air friction, or electrostatic forces)..

This leads to the truism that *a model is appropriate if and only if its predictions are accurate to within some pre-defined range*. In some sense, this means that a *model* can never be shown to be *false*; instead, it could merely be inappropriate for the given situation, due to aspects which are not taken into account. However, although the models themselves may not be falsifiable, the theory itself is still scientifically falsifiable *because the theory also specifies in what situations the generated models will be appropriate*. Falsifying a theory then amounts to determining that the models generated by that theory are not sufficiently predictive in the situations to which it is applied.

## **Computational Modelling**

Given the above considerations, mathematical modelling can be seen as a special, restricted case of computational modelling. Computational models are to be treated *in exactly the same way as* mathematical models, requiring the same theoretic considerations that were described above. Indeed, computational models may be best thought of as mathematical models *which are not analytic* (i.e. not formally solvable or able to be reduced to a closed-form solution). Due *solely* to their complexity, the standard mathematical short-cuts cannot be applied.

In particular, when developing computational models of cognitive processes, the convenience of calculus (or other mathematical proofs) is unavailable. This is a natural consequence of the greater complexity of cognitively active creatures, as opposed to the regular matter generally considered by physicists. This then means that situations like the three-body problem, as mentioned above, are the rule, rather than the exception. Fortunately, the solution is exactly the same: we convert the temporal behaviour of the model into small intervals of time, specifying what changes will occur over that time, and repeat the process until we have determined the model's prediction for long-term behaviour.

Since this iterative approach is best performed with a computer, it is termed *computational* modelling. Its history extends well before the development of the modern computer, and is strikingly similar to the task of traditional *human* computers, who were employed to do repetitive, detailed mathematical iterations to solve exactly this sort of mathematically intractable problem. More recently,

this sort of task is performed by digital computers, resulting in significantly faster and more accurate results.

Indeed, it is only the availability of such tools that makes computational modelling practical, as without such tools *it would be impossible to reliably determine the predictions of the model*. That is, digital computers can be seen as providing a new tool to science, which allows for model evaluation without requiring mathematical shortcuts. This is especially important because those mathematical shortcuts are only applicable to certain restricted circumstances, and thus constrain the sort of model that can have its predictions tested. We are now freed from that restriction, allowing for more detailed, and hopefully more accurate, models in domains which were not previously practical.

However, this is not to say that the scientific *technique* of modelling changes in any way. Instead, the exact process as occurred with mathematical models continues to be applicable to computational models. We still take the theories, customize them to create a model of the situation of interest, use the model to generate predictions, and then test those predictions. However, due to the expansion into more complex domains, care must be taken to ensure that the standard methodology is followed. The remainder of this paper details this process.

### **A Computational Model of a Cognitive System**

Theories developed in cognitive science are about systems described at a cognitive level, rather than at a physical level. This means that the theories should allow us to create models which predict the behaviour of living creatures, in the same manner as the theories of physics allow for the creation of models which predict, for example, the motion of planets. The test of such theories will be that it creates models which are predictive in certain well-defined situations. As in the previous example on planetary motion, the generated model is tested by making certain measurements pertaining to the situation, feeding those measurements into the model, and then evaluating the model to give predictions of other measurements. This is what we mean when we say a theory is *true*: it accurately predicts the outcomes that it claims it can. As the theories become more powerful, they apply to more situations, giving us a broader understanding and explanation of the behaviour of the whole creature being modelled.<sup>3</sup>

Just as the mathematical model of gravitation discussed previously can be seen as a model for taking masses and locations at one point in time and predicting the future locations, a cognitive computational model is thus a model for taking a set of internal and external influences on a creature and predicting its behaviour over time. Any particular model might restrict itself to certain sorts of influences, but this range of suitability should increase as the theories of cognition develop, as occurred with models of physics. However, instead of dealing with measurements such as *mass* or *distance*, there may be other sorts of values which will be more appropriate to cognitive models. Determining these points of comparison between the model and the phenomenon of interest is part of the responsibility of the theory itself.

### **Senses and Actions**

In any cognitive theory, it is also important to note the presence of the environment. Any situation where we want to predict the behaviour of a creature *in an environment* will require a representation of that environment within the model. The same is technically true of the mathematical model of

---

<sup>3</sup> The same mechanism can also be applied to modelling *parts* of a whole cognitive organism. The difficulty in this case is that most of the desired measurements are *internal* to the organism, and thus must be determined through highly indirect measures. However, the same principles do apply.

gravitation, but in that case the environment was assumed to be completely empty (or at least empty of any components that made a significant impact on the behaviour of the planets under consideration).

The mapping of the real environment to the modelled environment requires that the model generate inputs to the creature (senses), and adjust those senses based on the passage of time and the agent's outputs (actions). The precise methodology for defining senses, actions, and the effect of actions is purely defined by the theory. However, this definition is importantly constrained, in that there must be a defined mapping between the actions of the real creature and the actions in the model, and between the sensory environment of the real creature and that of the model. This is due to a methodological constraint: the predictions that we want from the model must be defined in terms of some observable behavioural response to different stimuli. Thus, in order to determine the predictions of the model, we need a method for converting the relevant aspects of the real environment to the modelled environment, and converting the actions of the modelled creature (its behaviour) into predictions of real behaviour.

The quality of this mapping varies depending on the situation being modelled. In some situations, it is appropriate to consider high-level models of this activity. The senses of the creature can be modelled by simply specifying what it sees in terms of objects, or just whether or not it can see a lever to press. However, such an approach limits the applicability of the model, as we then cannot use it for predicting how it will respond to novel objects, or to visual illusions, or to variations in how objects are presented. For these more complex situations, we need a lower-level mapping between the model senses and those of the real creature. In the extreme, this could be a complete visual image corresponding to the measurable activation of the organism's retina. Similar considerations are also necessary for the actions of the creature, which can vary from high-level actions such as “push the lever” to low-level actions such as contracting a particular muscle by a particular amount. Once this mapping is made, the remainder of the model can be specified.

## ***Evaluating Computational Models***

Any particular theory of cognition will thus allow for the generation of a computational model of organism behaviour in a particular situation. This behaviour can then be examined by running the computational model on a computer and recording the actions that it makes. This model behaviour can then be converted back into real-world behaviour for comparison to the actual behaviour in that situation. The accuracy of this prediction is thus the test of the theory.

### ***Parameters (Revisited)***

Given this view of modelling, the common problem of *parameter fitting* can be seen in a different light. The models produced by almost any theory will have a number of parameters, just as was true for the mathematical models. This leads to a common criticism of computational models; that they are merely the result of fine-tuned adjustment for a particular situation. Indeed it is certainly true that a sufficiently general model can be tweaked and adjusted to fit *any* desired set of data.

However, the same is also true for mathematical models. After all, mathematical models are merely a restricted set of computational models for which mathematical processing may be used to produce predictions, rather than requiring iterative computation. The reason that this seems to be a problem for computational models is due to confusion as to what claims are being made, and what the theory is about.

Just as the gravitational theory was useful for the first 120 years before the parameter  $G$  was

known, so too can computational theories which do not specify the parameters of their models. In these situations, the *application* of the theory requires some process whereby the parameter can be determined for the situation in question. Once this process, which usually involves using the model in some particular situation and finding the best-fitting parameter setting, is complete, the model can then be used to predict future aspects *of the behaviour of that particular cognitive agent in that particular situation*. That is, we perform parameter-fitting to create a model of this special case, and then can use that model to perform predictions. Importantly, it is this second stage which tests the truth of the theory. Merely finding a parameter setting which fits *does not inform us as to the veracity of the theory*.

A more detailed theory, however, may indicate particular values for certain parameters. Once a value for G was included within the theory of gravitation, it could be used in new situations *without the stage of first customizing the model*. This is clearly a desirable property for a cognitive theory as well. The theory could specify a particular value for a parameter, or even indicate a range of values, meaning that the model will be suitable *no matter where in that range the parameter is set*.<sup>4</sup>

It is also vital to note that parameters need not be merely numerical values. There is no reason why a theory might not treat an entire sub-module as a parameter. For example, the particular implementation of the world-model in the operant conditioning theory discussed above might be a parameter. In this case, a simple theory might say that we would have to 'fit' the model to a given situation (say, a rat pushing a lever) by finding a world-model component which gives a match to some aspects of the rat's behaviour. If the resulting model can then be used to predict other aspects of its behaviour, then we have a useful theory. However, a more developed model might specify one sort of world-model system which is to be used in a wide range of situations, or it might specify that any one of a family of models could be used and still produce accurate results. Indeed, it is always possible to build models with parameters which *function as if* they adjust between two different implementation systems, and this can also occur unexpectedly (see Sibley and Kello, 2004 for an example).

### **Processing (Revisited)**

The actual internal details of the models specified by a cognitive theory can be *any* algorithmic method. The only constraint is that it allow us to define certain values to constrain the model for a given situation, and allow us to determine certain values which are to be compared to the real-world measurements. This covers a vast variety of possible systems. Indeed, One of the major reasons for using computational models in the first place is that they allow us more flexibility than is found in mathematical models, which must be simple if they are to remain tractable.

Because of this, we are often in the situation of having *multiple* potential models (generated by different theories). These models may match to different ranges of situations, indicating different domains of utility for those theories. This is a typical situation from mathematical models in physics, as seen in the development of different theories of gravity, or the planetary models of Ptolemy and Copernicus. The initial, simpler, more restrictive models are a necessary first step towards the development of the broader models, and provide clear points of comparison.

However, there is a new situation possible due to the complexity of computational models. If a particular model does provide accurate predictions, it can be unclear as to whether or not there are aspects of the model *that could vary without affecting the accuracy of the predictions*. That is, the

---

<sup>4</sup> Merely saying that there will be a parameter value somewhere in some range for which the model will give accurate predictions puts us back to the original situation of needing to customize the model before applying it.

theory may specify a particular value for a parameter, but it may in fact be the case that the model would function well with a wider range of values for that parameter. If this occurs, then the theory *gives us a false sense of accuracy*. The theory may claim that a particular parameter should be 1.0, when it is just as accurate at a value of 0.1 or 10. A required part of testing a theory is then to show that the theory would *fail to predict accurately* if the parameters it specifies were outside of some range.<sup>5</sup> Interestingly, this implication of the model-based approach directly addresses one of the key concerns about model fitting raised in (Roberts & Pashler, 2000), which correctly questions the utility of merely indicating one particular set of parameter values for which a model is shown to fit.

More generally, this issue of varying values is also the the case for parameters which are not simple numbers. As mentioned in the previous section, whole sub-components of a model can be treated as parameters. The model of operant conditioning discussed earlier may predict well for a wide variety of action selection systems. If we only test one such system, then we are making the same error as only testing one particular value for a simple numerical parameter.

Unfortunately, this is a more difficult situation to deal with, as there is no way of systematically trying every possible alternate version of a particular module, or even every possible implementation that is in some sense 'close to' a given implementation. The space of possible versions is not one-dimensional. Certainly, a number of different versions can be tried, and some will result in accurate predictions, and some will not. Describing this range of suitable versions (and which versions are not suitable) *is as important as specifying what numerical parameter values are suitable and not suitable*.

### **Matching (Revisited)**

As was mentioned when discussing the gravitational model, the key question for determining the appropriateness of a model is whether its predictions match those of the real situation being modelled. This match is performed by measuring some aspect of the real world, and measuring some aspect of the model, and comparing the two. Of course, this match will never be *exact*. Instead, part of the theory involves specifying that the actual value will be within some range of the prediction (or possibly that it will be within some range a certain percentage of the time). A theory with a very wide range will thus be less generally useful than one with a more restrictive range.

This is true for both cognitive models and for the more familiar physical models. However, the behaviour of cognizing creatures tends to exhibit more variability than is normal in physics. Fortunately, there are various statistical measures which can be made to describe the variability, and this variability can also be predicted by the models.

However, the statistical tools required to do this are not currently well-known. Most statistical methods are concentrated on finding *statistically significant differences*, not *similarities*. For this reason, it is common in the literature to find the correlational measurement  $R^2$  used to give a numerical value for 'how close' a model matches the observed data. While this may be useful for determining that model A matches more closely than model B, it does not allow us conclude that a model with an  $R^2$  of 0.95 is in some sense *correct to within some range*, as we could with the mathematical model of gravitation.

There is, however, a relatively unknown statistical tool which is ideal for this situation. *Equivalence*

---

5 It is also possible that the model would give accurate predictions for some complicated set of parameter values, such as "any prime number", or "any value between 2-10 and 400-410". However, there will almost always be some range of values around any one suitable value which can be identified.

*testing* is a technique used in the evaluation of drug treatments to determine if a new, cheaper drug is as effective as some other drug, to within some pre-defined range. This is a modified version of the standard t-test, where instead of the traditional null hypothesis that the means of two groups are equal ( $\mu_r - \mu_m = 0$ ), the null hypothesis is that the difference between the means is *greater than some amount* ( $|\mu_r - \mu_m| > \theta$ ). The value of  $\theta$  defines the range of acceptable results. If we perform this statistical test, using  $\mu_r$  as the real data set, and  $\mu_m$  as the data from a given model, then a p-value less than 0.05 allows us to conclude with 95% certainty that the model and the real system do not differ by more than our threshold,  $\theta$ . This approach can also be applied to ensuring that other measures (such as standard deviation or skew) are also statistically indistinguishable. Furthermore, instead of setting the threshold and determining the p-value, we can instead set the p-value and determine the required threshold. This gives us a statistical measurement which has an intuitive interpretation: We are 95% certain that this model produces data that is no more than a certain value different from the real data.

As a final note, this sort of measure is suitable for *any aspect* of the behaviour of the cognitive system in question. A theory might be appropriate for predicting long-term trends, or short-term behaviour. It is also important to consider that what we call 'behaviour' can be any observable aspect of the system, and is not restricted to overt actions. For example, (Anderson et al, 2004) uses blood oxygenation levels within the brain as a measurable behaviour for prediction.

## **Further Implications**

This model-based approach to science sheds new light on a number of issues of scientific process. It is worth pointing out a few of these and describing how the modelling approach may give us a better sense of what is appropriate in this methodology.

## **Separating Hypothesis and Prediction**

The first thing to note is that we are *treating computational models as hypotheses*. That is, a particular computational model can be seen as a hypothesis that this computer program is predictive of some particular real cognitive behaviour. However, we are also noting that we must perform this complex simulation step in order to determine what the model predicts. In other words, the model is *opaque* in terms of its predictions: one cannot simply look at the model and know what it implies.

By identifying the computational model with the hypothesis, and by accepting the opaqueness of the model, we are in effect separating the hypothesis and the prediction. However, scientists are often more used to situations where the hypothesis is exactly the same as the prediction, or where the prediction is a trivial consequence of the hypothesis. In these simpler situations, we rely on the fact that if the prediction is true, then the hypothesis is true, and if the prediction is false, then the hypothesis is false.

However, in the case of computational modelling, we have a situation where the hypothesis (H) implies a prediction (P), and we then check the truth of our prediction. So, we have  $H \rightarrow P$ , and if we determine that P is false, then we can clearly also conclude that H is false (or, more accurately, that it is false to claim that the theory produces accurate models in this situation). However, if P is true (i.e. if the model is accurately predictive), then we have an interesting situation.

We clearly cannot simply conclude that H is true (i.e. that the model is 'true' or 'correct'). More precisely, we cannot conclude that our model is the only accurate method for producing predictions. This is because  $H \rightarrow P$  does not mean that  $\sim H \rightarrow \sim P$ . Finding one predictive model does not mean that other models are going to be less predictive. Certainly, it is a good first step, but it is quite possible that

other (possibly much simpler) models could also give a sufficiently similar result.<sup>6</sup>

This means that we should not follow the typical approach of testing a single hypothesis at a time. Instead, we generally require a multiple-models approach, with perhaps hundreds of models being tested. We have  $H_1 \rightarrow P_1$ ,  $H_2 \rightarrow P_2$ ,  $H_3 \rightarrow P_3$ , and so on. Only then are we in a situation to usefully identify one model (or one group of models) which gives sufficiently accurate predictions, while at the same time noting that other models do not. Without such evidence, we cannot be sure that there is something important about the structure of our model that produces the accurate predictions. After all, it could be that *any* reasonable model might produce data that is as accurate. It is incumbent on the modeller to disprove this possibility.

Most importantly, we cannot take the increasingly common approach of deciding upon one hypothesis that we think is the 'right' one, and running our experiment to confirm or deny this hypothesis. The experiment is meant to decide between competing hypotheses, and is not a situation where we should be *hoping* for one particular outcome (thus introducing potential biases). This is particularly a problem in computational modelling, where it is common for researchers to have particular modelling architectures that they favour.

Furthermore, since we cannot compare our models against all possible other models, we must perform our modelling in such a way that it is *easy for other researchers to do the exact same test using new models*. It is both highly likely and scientifically desirable that future researchers will want to continue to explore a given situation with new models, or make slight changes to the experimental situation to attempt to tease apart the results of currently equivalent models. One of the great advantages of computational models is that this level of sharing is possible.

### **Explaining versus Predicting**

A second issue to consider is in the fact that the modelling approach only mentions *prediction*, as opposed to *explanation*. One approach to this distinction is to deny that it is important. In *Science Without Laws*, Giere (1999) argues that science does not uncover universal 'True' Laws of Nature which explain how the world works. Instead, science is the discovery of *principles*. Giere's argument for the use of modelling entails that there is *only* prediction: explanation is simply a term for what happens when we have a model that is highly predictive in a wide variety of situations (Giere, 1988; Giere, 1999).

Principles, I suggest, should be understood as rules devised by humans to be used in building models to represent specific aspects of the natural world. Thus Newton's principles of mechanics are to be thought of as rules for the construction of models to represent mechanical systems, from comets to pendulums. They provide a *perspective* within which to understand mechanical motions. ... What one learns about the world is not general truths about the relationship between mass, force, and acceleration, but that the motions of a vast array of real-world systems can be successfully represented by models constructed according to Newton's principles of motion. (Giere, 1999, 94-95)

This view fits well within the framework described within this paper. However, it is not necessary for us to adopt this approach. Instead, we can insist that models which we are to consider as *explanations* are only those which do provide evidence of Pylyshyn's (1984) strong equivalence to the

---

<sup>6</sup> Indeed, in (Stewart, West, & Coplan, 2004) we found that some predictions of our model of children forming friendships could also be generated by a completely random model.

real cognitive system. This translates to a requirement of being able to predict such ancillary information as reaction times, grain sizes appropriate for cognitive penetrability, and perhaps even reacting in similar ways to physical disruptions such as lesions.

Interestingly, models which do have all of these similarities of prediction and structure can also be seen as *real patterns* in Dennett's (1991) sense. That is, the components of models which result in effective predictions can be interpreted in the same manner as beliefs and desires would be interpreted in folk psychology, or as a glider would be interpreted in Conway's Game of Life. As Dennett points out, whether this should be seen as realism or instrumentalism is a matter of definition.

It is also worth noting that without a rigorous definition of what is to count as an explanation, we may fall back on misleading intuitions. Lombrozo (2005) showed that people are quite willing to accept explanations based on adaptationism, even though they themselves have little understanding of the basic components of how an adaptationist explanation should work. Ofttimes merely labelling something as an explanation is sufficient for us to believe we have an explanation, rather than basing such a decision on the rigorous predictive abilities of the model.

The key factor here is that the more a model predicts (to a useful degree of accuracy), the more confidence we should have in its overall explanatory power. In particular, what we want are models which predict *a wide variety of behaviours*. Even more convincing are models which are developed for one domain and then expanded into new domains. A successful expansion of this sort should be an indication that the model is genuinely predictively useful, and thus we should have more confidence in its veracity.

### ***Behaviourally Identical Models***

A further quandary arises when we find two (or more) models *of completely different types* which are equally good at predicting the real-world data. This is clearly theoretically possible; given any computer program, it is always possible to re-write it into a new program which gives *exactly* the same outputs for the same inputs, and yet has entirely different source code (Moore, 1956). It is not even theoretically possible to consistently detect this sort of convergence; we cannot, in general, know that two programs will give identical inputs and outputs without trying all of the infinite number of inputs, which is certainly not a practical methodology. So what should we do when we find two models wildly different in implementation, and yet identical in behaviour?

This is, of course, the same problem as in all of science when there are two competing theories, and they both give the same predictions. Some may argue that the 'simpler' model (in some, vaguely defined, sense) is the one that should be used, although this seems to be inconsistent with the actual practice of science. What actually tends to happen is that both models stay in use, with one usually significantly dominating in presence over the other. The deciding factor may then be *which model is more conducive to further exploration of the phenomena*.

There is an interesting analogue to this situation occurring within the physical sciences is from the field of quantum physics. Both the Copenhagen Interpretation and the Bohm Interpretation give *exactly* the same predictions (for all situations where the Copenhagen Interpretation is unambiguous). However, the Copenhagen Interpretation (that 'observing' quantum events 'collapses the waveforms') is the dominant one, both among quantum physicists and within popular media. For example, the Schrodinger's Cat thought experiment (that if you kill a cat based on the outcome of a quantum event, you end up having to treat the cat as both alive and dead at the same time until you actually observe it)

is only seen that way within the Copenhagen interpretation. In contrast, the Bohm Interpretation (that there are a large number of hidden, non-local variables that we can see the results of but not actually measure) leads to no such situations, but has not received wide-spread popularity among either scientists or the public.

There seem to be a number of *sociological* reasons for this difference, including meme-spreading or simply the attraction of strange ideas. However, there is also a very good *scientific* reason why one might pick one model over another, in the absence of any distinguishing data. One model may be better for *thinking about the problem*. This is, it may be easier to work with, or tend to more readily lead to interesting or useful predictions.

With this in mind, finding multiple highly different models which give the same predictions is not a problem for computational modelling. We keep them all, and can continue to use them at our convenience. If future information allows us to decide between them, then we can do that; in the meantime, there should be no pressure to determine which is the *real* model. Indeed, most quantum physicists regard worrying about these multiple interpretations to be a waste of time.

## **Conclusion**

The methodology presented here can be applied to any example of computational modelling within science, but it is particularly appropriate for situations involving the creation of artificial cognitive systems. Certainly, such systems can be studied without such an approach, and many artificial life simulations are worth studying in their own right. However, if we are to use models as scientific explanations of the behaviour of real living creatures, we must carefully consider how we are to relate the models to the real-world situation being modelled.

This approach relies on us viewing theories to be instructions or principles which allow us to examine a particular situation and construct a model which will provide accurate predictions about that situation. As we develop the models, we measure their predictive accuracy (using tools such as equivalence testing), and we identify the domain of situations for which they are usefully predictive. Importantly, this does *not* mean that all aspects of the model must match to aspects of the real situation. We do not have to ensure that the models are biologically plausible, or that particular components of the model work in exactly the same manner as we believe the real creature does. After all, physicists know that Newton's law of gravity is not an exact match to the real world, as it cannot take into account modern relativity theory. Instead, physicists have developed strong guidelines for identifying *what situations the theory can be applied in anyway*, and for those situations the model does give us remarkably precise predictions. It is this approach that should be adopted for models of cognitive systems as well.

To summarize the approach, a computational theory about a real cognitive system must specify exactly how to define a model for a given situation. This specification must also include specifications for each parameter and component within the model, or a methodology for determining these aspects for the particular situation. It must also identify the measurable aspects of the real system for which it is predictive, and to what degree. These predictions should not be merely concerned with the mean observed values, but also other distributional measures. Furthermore, any such model needs to be accompanied by evidence that using values outside of the ones indicated by the theory, or using alternate models, leads to less accurate predictions. Clearly this evidence can never be complete, so we must ensure that it is as easy as possible for others to investigate other possible models for that situation.

## **References**

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review* 111(4). 1036-1060.
- Dennett, D. (1991) *Real Patterns*. *Journal of Philosophy*, 88(1), 27-51.
- Einstein, A. and Infeld, L. (1938). *The Evolution of Physics* (New-York: Simon and Schuster), 1961.
- Giere, R. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press
- Giere, R. (1999). *Science without Laws*. Chicago: University of Chicago Press.
- Lombrozo, T. (2005). Why Adaptionist Explanations are So Seductive. 27th Annual Conference of the Cognitive Science Society.
- Moore, E. F. (1956). Gedanken-experiments on machines. In *Automata Studies*, ed. C. E. Shannon, J. McCarthy. Princeton, NJ: Princeton
- Pylyshyn, Z.W. (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.
- Roberts, S., & Pashler, H. (2000). How Persuasive is a Good Fit? A Comment on Theory Testing. *Psychological Review* 107(2) 358-367.
- Sibley, D. E. & Kello, C. T. (2004). Computational explorations of double dissociations: Modes of processing instead of components of processing. *Cognitive Systems Research*, 6, 61-69.