Validating biorobotic models

Barbara Webb

Institute for Perception, Action and Behaviour
School of Informatics
University of Edinburgh
JCMB Kings Buildings
Mayfield Rd
Edinburgh EH9 3JZ
United Kingdom

E-mail: bwebb@inf.ed.ac.uk

Abstract.

Some issues in neuroscience can be addressed by building robot models of biological sensorimotor systems. What we can conclude from building models or simulations, however, is determined by a number of factors in addition to the central hypothesis we intend to test. These include the way in which the hypothesis is represented and implemented in simulation, how the simulation output is intepreted, how it is compared to the behaviour of the biological system, and the conditions under which it is tested. These issues will be illustrated by discussing a series of robot models of cricket phonotaxis behaviour.

1. Introduction

Many researchers are now engaged in building artificial systems intended to model some aspect of biology, such as the simulation of neural systems to emulate some of the processing capabilities seen in natural neural systems. Although in many cases the systems are constructed and tested in software, there is also much interest in embedding neural systems in hardware to evaluate their capabilities in the real world. In cases where the hardware used is a robot, and the aim of the evaluation is to explain the behaviour of some biological system (not just to mimic it), this modelling approach has been termed 'biorobotics'. However, the practice of modelling is not always accompanied by good understanding of the methodological issues that arise in simulation, experimentation and validation. Such understanding can be a useful guide to both building and interpreting the results of models.

In (Webb 2001) I presented a framework for thinking about modelling in general and biorobotics in particular. This schema is illustrated in figure 1. We choose a *target* system to represent some interesting problem in the world, e.g. the phonotaxis (sound localising) system of crickets as an example of sensorimotor control. We then *hypothesise* a causal mechanism to explain how this system functions. At least some of the ideas for this mechanism may come from some pre-existing *source*, e.g. by drawing an analogy between the cricket's tendency turn to the side of louder sound and the simple vehicle controllers described in (Braitenberg 1984) which speed up or slow down a motor proportionally to the strength of a signal on the same or opposite side. We then implement this mechanism in a *simulation* with the aid of some *technology*, e.g. a mix of robot hardware and microprocessor software‡. The implemented system will produce some behaviour, and can be tested under circumstances that represent the original problem e.g. can it find one among many sound sources? The behaviour of the simulated system is interpreted and compared to the corresponding animal behaviour, e.g. by statistical comparison of the choices it makes between sounds (Webb 1995).

The obvious aim in carrying out this procedure is to test whether our hypothesis about the mechanism was correct (see Webb et al, forthcoming, for discussion of other possible aims). Note that this applies even when our intention is not so much to understand any specific existing target system but rather to solve a given problem. E.g. our *target* might be to find a good method for generalisation in a sound recognition system, and the mechanism proposed may have its *source* in biology. Nevertheless these is still an identifiable process of forming a hypothesis about a possible mechanism, implementating it in some simulation system, generating behaviour, and comparing the behaviour against our original target specification. Our prime interest is to show that the hypothesised mechanism can actually perform the target capability; and often

‡ In most of what follows, I am using 'simulation' to include robots, i.e. they are simulations of biology, not real biological systems. There are some differences between implementing a model purely in software, and implementing it in a robot that behaves in the real world, but the differences are not essential to the points I wish to make in this paper

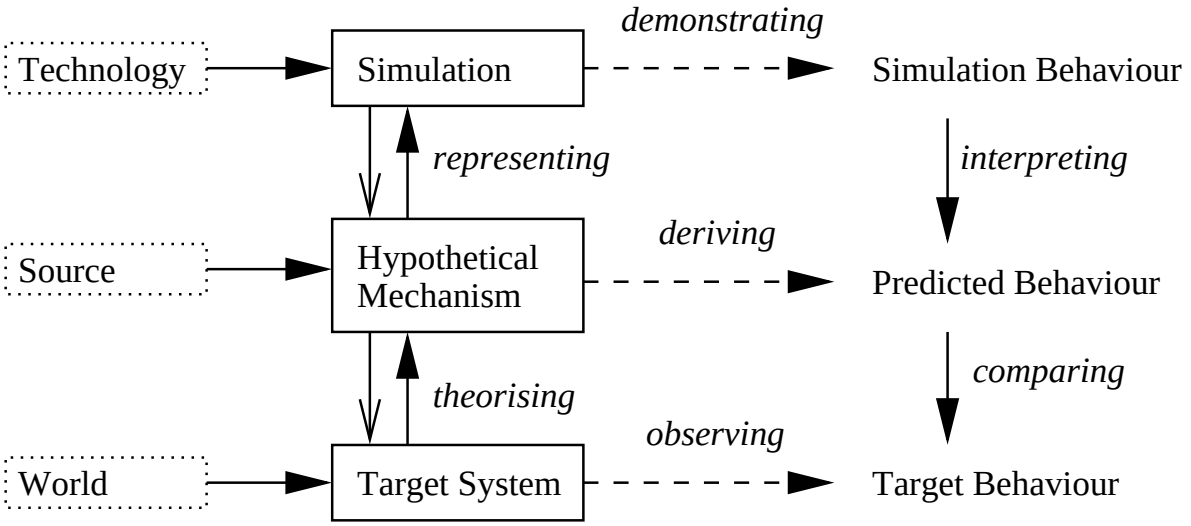

Figure 1. A schematic outline of the process of building a simulation to explain some target behaviour

subsequently to investigate how it performs under a range of conditions not necessarily foreseen; in the case of a scientific model, to make new predictions about the original target. If it works, we take the hypothesis to be supported, and if it does not, we conclude the hypothesis is incorrect.

A central problem, however, is that drawing any such conclusions from this process is not straightforward. First, if the comparison of behaviour appears successful (the simulation output matches the target output) we cannot simply conclude our hypothesis about the biological system is correct§. It is always possible that some other, different mechanism could account just as well for the data. This issue is sometimes called 'weak underdeterminism': any finite set of evidence could be entailed from indefinitely many alternative theories (Laudan 1998). ‖

If we do have two alternative simulations that both produce the right output we may nevertheless distinguish them on other grounds (simplicity, generality, the fit with other theories and so on); in other words we might argue that we currently have the 'best explanation' (Harman 1965). But we cannot be certain. One strategy (Lipton 2004) could be to attempt a Bayesian estimate of how well the hypothesis is supported by the match between the simulation and real system data:

$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$.

This formulation makes explicit several issues (although obtaining true probability estimates for them is likely to be impossible). The prior plausibility of our hypothesis,

§ Even concluding that we have at least found one solution to the problem must also be subject to some reservations, as will be discussed below

‖ 'Strong underdeterminism' is the claim, made by some philosophers of science, that there can never be enough evidence to decide between theories: this is rather trivially true if we allow ad hoc hypotheses such as "any evidence contradicting my theory was created by demons"; but more interestingly might be true when the theory itself entails the impossibility of gathering the required evidence.

$P(hypothesis)$, will affect our estimate. For example, if the simulation implements a hypothesis that has previously had wide experimental support, we might draw stronger conclusions than if it suggests a completely novel and somewhat unlikely mechanism. Similarly, if fitting the data is simply an exercise in parameterising polynomials, we are unlikely to consider the resulting equations as a plausible hypothesis for the mechanism, despite the perfect match of outputs (although they may be revealing about possible mechanisms, particularly if the polynomials are of low order). The prior likelihood of the data, $P(data)$, also affects our estimate because, for some data, it might be easy to think of many ways it could be produced. For example, simply being able to follow sound would only weakly support the argument that the robot works in the same way as the cricket, whereas following only cricket sounds and producing similar patterns of approach is less probable, and hence more supportive.

Finally, considering the likelihood of the data given the hypothesis, $P(data|hypothesis)$ makes explicit the fact that the output does not follow directly from the hypothesis itself, but depends on the rest of the simulation process in figure 1. The fact that the output matches might be due to some particularity of the implementation, or the conditions under which the system was tested, or the manner of interpreting the results, or due to flexibility in the matching criteria. For example, the cricket robot uses wheels, not legs; possibly the control mechanism that produces phonotaxis on the wheeled robot would not work on a walking robot. Because the robot has wheels, it is tested on flat, hard surfaces, which do not resemble the normal conditions in which crickets do phonotaxis. In the early version of the cricket robot described in (Webb 1995) a slower processor meant the the sound stimulus was slowed down, but it was assumed a scaling factor was sufficient to treat the robot tracks as comparable to cricket tracks - but this provides a free parameter that allows the 'match' to be tuned to fit the data. The tracks may appear qualitatively similar, e.g. with the robot making the same choice as the cricket, but not be a close quantitative match. Thus caution must be exercised in drawing strong conclusions from an apparently successful comparison between simulation and target behaviour.

The opposite case, i.e. failure of the simulation output to match target output, might on the face of it seem undesirable. However it has a logical advantage over confirmation: although we can never be sure a theory is true (as discussed above) we can reject a theory that has been falsified ((Popper 1968) although see below). In practice, modellers often find that more is learnt from the failure of a simulation than from its success. This idea underlies one strategy in modelling which is to start with very simple mechanisms, which are almost certain to fail, and incrementally elaborate the hypotheses in the light of how the simple versions fail to account for the data.

The problem remains, however, that failure to replicate may not necessarily be a result of having the wrong hypothesis or proposing a mechanism that doesn't work. Instead, it could be because the implementation did not represent the hypothesis correctly, or the testing conditions were inappropriate, or that the interpretation was incorrect, or the match involved the wrong comparison, or even that the experimental

observations on the biological system were flawed. In other words, modellers need to be convinced that none of these factors apply before they can conclude that the hypothesis itself is falsified. In practice, failure to produce the desired results rarely leads modellers to directly reject their hypothesis: more commonly they tinker with one of the other factors. The equivalent problem is often discussed in the philosophy of science as the Duhem-Quine thesis (Harding 1976). This states that every actual experiment is dependent on a large number of assumptions additional to the hypothesis in question: assuming that the experimental apparatus works correctly (and the science that underlies the design of the apparatus is correct); assuming that a large number of external conditions (the weather, the colour of the experimenter's t-shirt) make no difference to the outcome ("ceteris paribus" (Cartwright 1983)); and so on. Failure to get the result predicted by the hypothesis can always be blamed on one of these auxiliary assumptions, rather than an incorrect hypothesis.

This flexibility or indeterminancy might seem rather negative — can we never decide anything about the truth of a hypothesis by modelling or experimentation?. A more positive way of looking at it is that our model-building can be enriched by explicit consideration of what kinds of changes can be made to each step in the process in figure 1. In what follows, I will use the example of our work on robot models of cricket phonotaxis to illustrate some possibilities:

- Changing the way in which the hypothesis is represented in the simulation.
- Changing the conditions under which the simulation behaviour is demonstrated.
- Changing the interpretation of the behaviour of the simulation.
- Changing the criteria for comparing the behaviour to the target.
- Changing the conditions for observations on the target system.
- Changing the hypothesis itself.

2. Changing the representation

One distinctive feature of biorobotics is that it involves moving from purely software simulations to simulations that involve at least some hardware as part of the technology for implementation. This can have both advantages and disadvantages when it comes to the problem of whether the implementation correctly represents the hypothesis. For example, the difficulty of accurately modelling the physics of water and land-based locomotion is a clear motivation for building physical versions of salamander models in work by (Ijspeert et al. 2005). But constraints on what can be built mean that the size is increased and number of segments decreased in comparison to a real salamander. Moving between different implementations of a model, such as both software and hardware versions (as has been done for the salamander model), can be a useful strategy in these circumstances. It allows the central hypothesis to be tested under different, complementary, sets of assumptions, thus reducing the likelihood that it is the representation, rather than the hypothesis, that is determining the output.

It is worth keeping in mind that validating the implementation may include the simple issue of whether the software or hardware actually does what you intend it to do. Some bugs may be hard to spot, especially if their effect is to (erroneously) produce the correct behaviour. Within software programming there are approaches that aim to improve (even formalise) code verification, but these are rarely applicable to the kinds of programs used in neural modelling. The same problems may apply to hardware - does a component actually perform according to its design specification? Is the implementation supported by sound mechanical or electrical theory, or just engineering experience? Some issues, such as communication speeds, can fall on the software/hardware boundary.

Any simulation implementation can potentially be changed in many ways. Having built a simulation, such as the first robot cricket (Webb 1995) that seems to reproduce the cricket behaviour, why should we go on to build further models of the same behaviour? There are several motivations. One is that by including more constraints, we can try to improve support for the possibility that the mechanism is accurate to real biology, and not just coincidentally able to produce similar output. Another is that we may want to account for lower levels of the mechanism; in the case of sensorimotor behaviour we might want to explore the neural implementation of a particular control mechanism. Or we might want to account for more of the known details about the system. Work on the robot cricket can illustrate changing the level, the detail and the accuracy, as follows.

The first implementation of a controller to mimick the cricket behaviour (Webb 1995) was in the form of a Braitenberg-like algorithm (figure 2a). The auditory input on each side, after passing through a phase cancelling process that allows sound of a specific frequency to be lateralised, was integrated over time, and whichever side first reached a threshold would initiate a turn in the relevant direction, while suppressing the opposite side. The first layer of leaky summation acts as a lowpass filter, and the second, acting on sound onsets, as a highpass filter for the pattern of repeated sound bursts that make up the typical cricket song. This algorithm could produce an interesting range of behaviours in a robot, including robust localisation of the sound, and discrimination between different sound patterns. The mechanism is thus a plausible hypothesis for the behaviour, but it is not clear whether the cricket could actually work this way. One way to test the latter issue is to try re-implementing the algorithm in terms of spiking neural units, and seeing whether it can still be made to work (Webb & Scutt 2000).

I would describe this reimplementation (shown in figure 2B) as a change in level, as it alters the base units in which the mechanism is implemented. This can be differentiated from increasing detail, because the initial circuit used was not closely based on known details of the internal neural connections of the insect; in fact it contradicted some of these, while still being quite informative about how certain apparently complex behaviour (e.g. choosing between similar sounds) could emerge from a small circuit of neurons. However our next model (Reeve & Webb 2003) used a network that included more of the actual structural details about the internal network in crickets (figure 2C: such as the cross-inhibition occuring at the thoracic level (the 'ON1' in (Horseman &

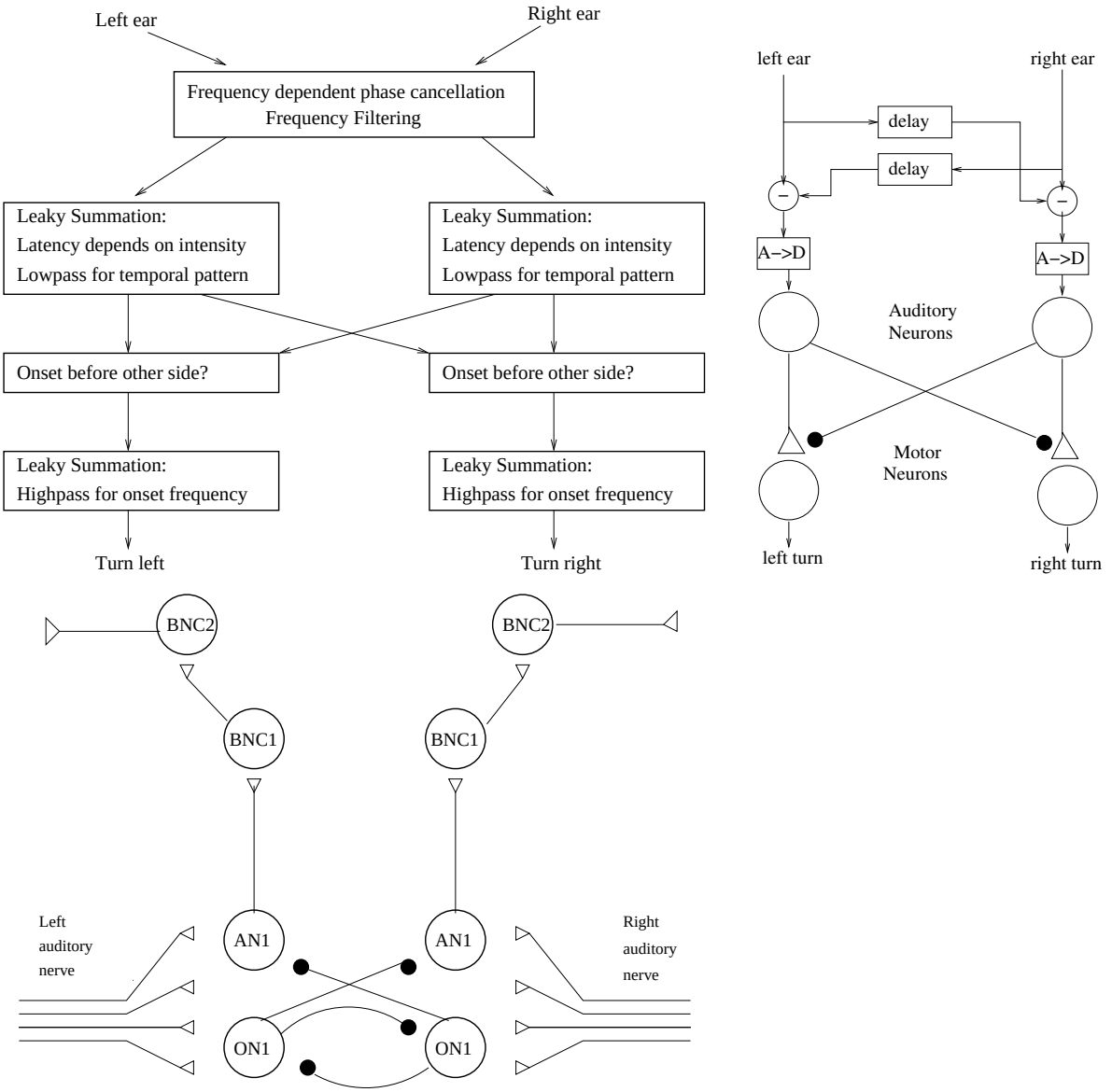

Figure 2. Successive models of cricket phonotaxis implemented on a robot. A: a Braitenberg style algorithm B: the same principle implemented in spiking neurons C: A spiking neuron model that represents known neural connections in the cricket in more detail.

Huber 1994)) and the two stages of processing in brain neurons 'BN1' and 'BN2' that have been associated with temporal filtering of the sound pattern (Schildberger 1984). This allowed us to test how well the mechanism account for neurophysiological as well as behavioural data. This new model also introduced more accurate modelling of the neurons and synapses. Without changing the level - i.e. the model still described membrane potential change in single compartment neurons - the implementation was revised to introduce exponential rather than linear decays, and to have synaptic action correspond to membrane conductance change rather than treated as charge-dumping (Koch 1999). One reason for doing so was to impose more realistic constraints on

parameter tuning (see below).

In this case the changes to the representation proceeded from simple to more complex models, and from higher to lower levels. However, model revision need not necessarily progress in this direction. It can be very productive to determine that some complex component of a model can be replaced by a simpler version without significant consequence for the hypothesis under test - i.e. to carry out a process of 'model reduction'. A good example comes from the previously mentioned work on models of lamprey swimming by (Ijspeert et al. 2005). Earlier models used neural circuits as central pattern generators, but later models represented these elements with simple equations for non-linear oscillators, as this was found not to significantly affect the critical questions about the coupling of oscillators that were under investigation.

An issue already hinted at in this discussion is that if the hypothesised mechanism has a number of components, then it may be necessary to consider validation of the implementation of each of the components as well as of the system as a whole. For example, we may want to show that specific simulated neurons, e.g. 'AN1' behave in the same way as the corresponding cricket neurons. But this effectively spawns a whole set of submodels, each of which can be described by its own version of figure 1. AN1 is now the target; we have a hypothesis about how it comes to have its observed behaviour; we have an implementation in software; and we look at the simulation output, comparing the spiking patterns to the real neuron. Similarly, hardware components of the model may also need to be validated, independently of the rest of the model, to demonstrate that they have the right capabilities. Recursively, the same set of problems arise for each individual component - can we be sure the implementation is correct, have we tested it under the right conditions, are we correctly interpreting the behaviour, does it match well enough? - before we can decide that our hypothesis about the mechanism for this component was correct. Note however that in this process we usually bottom out in the 'base units' or lowest level of the model. Here we no longer propose any explicit mechanism, but simply require the component to have the right input-output relations, using whatever mechanism is convenient e.g. in the robot cricket, using a Poisson distribution to generate a spike train proportional to input stimulus amplitude rather than modelling mechanisms of auditory transduction (Reeve & Webb 2003).

To summarise the main points made in this section:

- Building several models using different technologies is one useful way to guard against false conclusions resulting from a particular implementation.
- Don't assume that because your simulation produced the 'right' result that it contains no bugs.
- Going down (or up) a level, including (or excluding) detail, and increasing (or decreasing) accurary, may be useful for answering different kinds of questions about a system. The question should come first: it is rarely worth the effort of making a simulation more complex if you do not know what could be learnt by doing so; nor is it worth building a simulation too simple to represent the problems of interest.

- There is no 'best' level for modelling, but there must be some level at which we are no longer concerned whether the components produce the output in the same way as biology, otherwise our model validation will regress to proving quantum mechanics.

3. Changing the demonstration conditions

Another aspect of the modelling process that can be altered are the experimental conditions under which the simulation is tested. In this case I am not referring to the specific variables being manipulated in testing - such as varying the patterns of sound stimulus as described in section 5 - but rather the general conditions (sometimes called the boundary conditions (Tamburrini & Datteri 2005)) within which the experiment occurs - such as the use of specific sound-card and speaker to produce the sounds. These are usually assumed not to affect the outcome of the experiment, although it is also accepted that there is some range within which the conditions need to fall e.g. a certain quality of sound reproduction is necessary. Typically the conditions under which the simulation behaviour is produced are not identical to the conditions under which the target behaviour is produced, reflecting different assumptions about what might or might not affect the results. If the robot lacks a visual system, then experiments on sound localisation do not need to be done in the dark, as they usually are for crickets.

As an example we can consider the acoustic environment that has been used in testing the behaviour of robotic and real crickets. In most experimental paradigms, crickets are tested in anechoic chambers, or at least in surroundings designed to minimise sound reflection and any background noise. By contrast, most of the robot experiments have been conducted in an ordinary robot lab environment, specifically to demonstrate the robustness to reflected and extraneous sound; with the assumption that better sound controls could only improve the behaviour observed. However the ability to directly compare the paths produced by the robots and crickets might be limited by this difference in conditions. More recently, the robot has been tested in more naturalistic sound conditions, i.e. outdoors (Horchler et al. 2004). While this allowed the demonstration that the implemented mechanism continues to work for locating sound in this environment, the additional constraints imposed on the robot mechanics to enable it to move over natural terrain reduced the possibility of obtaining similar paths to the cricket, particularly because it increased the robot's turning circle. Moreover, comparable data on cricket phonotaxis paths under natural conditions are not available for comparison. These considerations have led us to build an arena in which we can test both insects and robots under the same conditions.

Ideally, for the biorobotic approach, the conditions of testing would be identical; the robot would be tested in the same environment with the same stimuli as the animal. An example is the 'robolobster' (Grasso et al. 2000), which was built to run in the same test tank as that used for experiments on chemotaxis in the lobster, and has also been tested in real sea conditions (Grasso, personal communication). This does not rule out

the possibility that the conditions are affecting the behaviour in some way that limits the testing of the hypothesis, but helps ensure that similar effects are occuring for both systems, so differences in their behaviour are not due to differences in these conditions. However, it should be noted that even when using the same environment for testing, it is usually not the case that all aspects of the agent-environment interaction are identical. In the robolobster example, it is detecting conductivity in a saline plume rather than the food chemical plume tracked by lobsters; similarity in the dispersal characteristics had to be verified (Grasso et al. 2000). Even when using identical stimuli and sensors, as in the case of moth chemotaxis using a robot fitted with real moth antennae (Kanzaki et al. 2005) the flow of the plume around the moth body might not be identical to that around the robot.

Another important issue is that for real biological systems, there may well be learning or adaptation of the system in response to the environmental and experimental conditions. We can try to include adaptive mechanisms in our simulation, but if the conditions of testing the simulation differ, this will tend to amplify any differences. It is a well-known problem that evolutionary simulations tend to adapt to exploit the specific agent-environment interactions in which they operate, and in the process may 'finesse' the problem in a way that tells us nothing interesting about the possible biological mechanism.

To summarise:

- As for different implementations, it can be useful to test the simulation under different conditions, so that the output is less likely to depend on any specific conditions.
- It is a good aim to make the conditions for testing as similar to the biological system as possible, and a robotic (rather than computer simulated) implementation may help for this, but it is very difficult to ensure the conditions are really identical.
- Adaptive mechanisms in animals and robots will tend to amplify any difference in the testing conditions.

4. Changing the intepretation

In interpreting the output of the model, we are specifying a mapping between the simulation performance and the target system performance. It can at times be too easy for mere labelling of the output (e.g. calling a particular state in an animat 'hunger level') to convince us of identity (i.e. to assume this is equivalent to food deprivation in an animal experiment) and to fail to confirm the grounds for the mapping. If the mapping is in fact rather weak, then the interpretation of the behaviour may be distorted. An example is describing a simulated behaviour as 'chemotaxis' when there is not only no real chemicals involved, but the dispersion of the 'chemical' is modelled as a simple gradient, which is highly unrepresentative of the nature of real chemical plumes. Interpreting the behaviour as chemotaxis is misleading if this is meant to support the conclusion that the simulated mechanism can explain real chemotaxis in biology.

Simulations of neurobiology often do not produce real behavioural outputs but intepret some output (e.g. rate of firing of some neuron, or the pattern across a neural population) as corresponding to some behaviour. This can be problematic when there is an assumption that intended actions directly correspond to actual actions, ignoring the effects that real embodiment might have on the output. It may in fact not be possible to get from the simulation to the correct behaviour; the body constraints on input and output may require a very different mechanism. An interesting example is the 'place cell' recognition system in the rat. There are a number of robot models that are able to produce responses in a neural layer that can be interpreted as 'place cells', e.g. (Mataric 1990, Burgess et al. 1997, Arleo & Gerstner 2000, Hafner 2005). Some of these identifications are more tenuous than others: if the criteria is only that the unit responds when the robot is in particular location, it seems that many possible mechanisms could explain the response. Stronger constraints on the interpretation need to be introduced if the models are to be seriously considered as hypotheses for the rat.

In the case of biorobotics it might appear that the intepretation step is straightforward - the robot is producing real behaviour that can be compared in a direct way to animal behaviour, e.g. does the path of the animal match the path of the robot? In the case of place cells, unfortunately, very little is known about how place cells are actually used by the rat to control its behaviour. In other examples, such as crickets moving towards sound, there may be better information available. But, nevertheless, any comparison involves some abstraction, as the two systems may not be directly comparable in speed, size etc. and almost never have directly equivalent locomotion systems. For example, it was noted in research on the sahabot, which models desert ant navigation and was tested in the same environment that:

> Athough the distances covered in the excursions of ants and in the robot experiments are in the same range...it is difficult to compare the homing precision of these agents, since both their size and their method of propulsion are completely different (Lambrinos et al. 2000)

The modeller may consequently introduce mapping factors that improve the match without actually being explicitly part of the hypothesis. There is a risk that if these mapping factors are used flexibly, they may be doing as much work in producing a successful match to the target behaviour as the hypothesised mechanism itself.

Another important way in which comparison can be limited has to do with the substantial inequality in the kinds of experiments that can be done, and the kind of data that can be gathered, for simulated and real neural systems. In the neural model of the cricket, we can measure the activity of every neuron and synapse, and correlate these directly with behaviour. We can also selectively manipulate the activity and connectivity of every neuron and examine their effect on the behaviour. Few such experiments are practically possible for the cricket. As a result, much model building must be regarded as exploratory in nature and not subject to the normal ideas of confirmation or validation.

In summary, the following should be kept in mind when interpreting the output of

simulation:

- Be wary of labelling outputs as though they are real behaviours, and then interpreting the results according to the labels rather than the real output.
- Also beware of convenient parameters that allow the simulation behaviour to be arbitrarily tuned to provide an apparent quantitative match with the biological system.
- For validation to be possible, a model must be focussed towards producing data for which the equivalent biological data exists, or is viable to obtain.

5. Changing the comparison

The issue of how the behaviour of the model will be compared to the behaviour of the target system is partly an issue of intepretation (as described above) but for a given interpretation there remains the issue of what degree of similarity is considered sufficient. A closely related problem is that of parameter tuning in models. It is well recognised that, given enough free parameters in our model, we can obtain an arbitrarily close match to any given data set. As a consequence, it is sometimes argued, modelling always succeeds, and hence is not informative. Or at least, that a model to be useful must predict new data, not just account for existing data.

An example for the cricket robot is the mechanism for recognition. Female crickets show a band-pass preference for the syllable rate in male cricket songs (each syllable is a pulse of sound lasting around 20ms, produced as the male cricket closes its wings). In our model (Reeve & Webb 2003) this preference is a result of the dynamic properties of synapses. Referring to figure 2C, from the auditory nerve to AN1, the time constant of integration is long enough to obscure the temporal pattern at fast syllable rates (acting as a lowpass filter). From AN1 to BN1, depression in the synapse results in firing of BN1 only at syllable onsets. From BN1 to BN2, the recovery rate requires a minimum rate for onsets to bring BN2 above threshold (acting as a highpass filter). The result is a bandpass preference; and the parameters can be tuned to make this preference similar to that seen in the cricket. However an alternative model, proposed by (Nabatiyan et al. 2003) is based on the observation that the onsets of syllables cause peaks in the firing rate of ON1 (and it is assumed, AN1), and that the rate of peak occurence can produce the preference tuning curve. In this case tuning of the model consists in setting a peak rate threshold to detect these onsets.

These models both account for the cricket's response to test stimuli in which the syllable rate of the song is gradually varied. But they could be better compared by looking at a greater range of stimulus paradigms. One useful way to more thoroughly characterise the cricket's behaviour is to consider how it responds to songs varied systematically on two dimensions, the pulse length and the length of the gap between pulses (figure 3). If these are varied with $pulse = gap$, we have the standard experiment in which different syllable rates are compared. If varied with $pulse + gap = constant$,

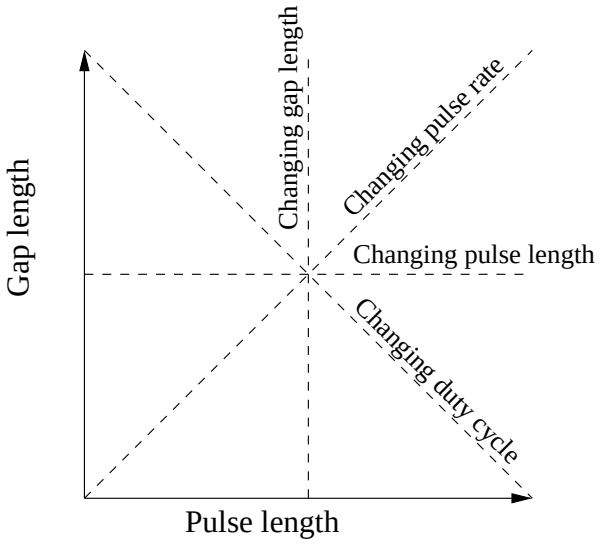

Figure 3. The space of stimuli for testing the pattern filtering properties of different networks

we are testing sensitivity to duty cycle. Or we can examine what happens with varing pulse and constant length gap, or varying gap and constant length pulse. We can also look at the response with different sound amplitudes. The results are shown for the two models in figures 4 and 5, and can be compared to the cricket results in figure 6 (from Hedwig and Poulet, personal communication). As can be seen, the first model provides a closer pattern of responses to the cricket for the duty cycle and pulse length; the second model has a relatively flat response as these are varied, whereas the first model and the cricket show bandpass selectivity. Interestingly, both models predict that in varying the pulse length, a stimulus with a shorter length than the normal cricket song should actually be more attractive than real songs. However the results of testing the cricket with the corresponding stimuli are not yet known. The second model also turns out to be highly sensitive to setting of the threshold parameter (equivalent to requiring the cricket nervous system to be sensitive to very small differences in the duration of gaps between spikes) whereas the first model is more robust.

Some strategies to consider:

- It can sometimes be useful to demonstrate what the model cannot do. A model that can potentially be fitted to any data set is not a useful one - in the same way as a theory that makes no falsifiable predictions is a weak one.
- Try to constrain the model parameters, as far as possible, on grounds other than fitting them to the data, for example by attempting to directly measure the relevant values for the system, such as neural time constants.
- Try varying the experimental stimuli outside the initial paradigm. The results may suggest critical experiments for biology.

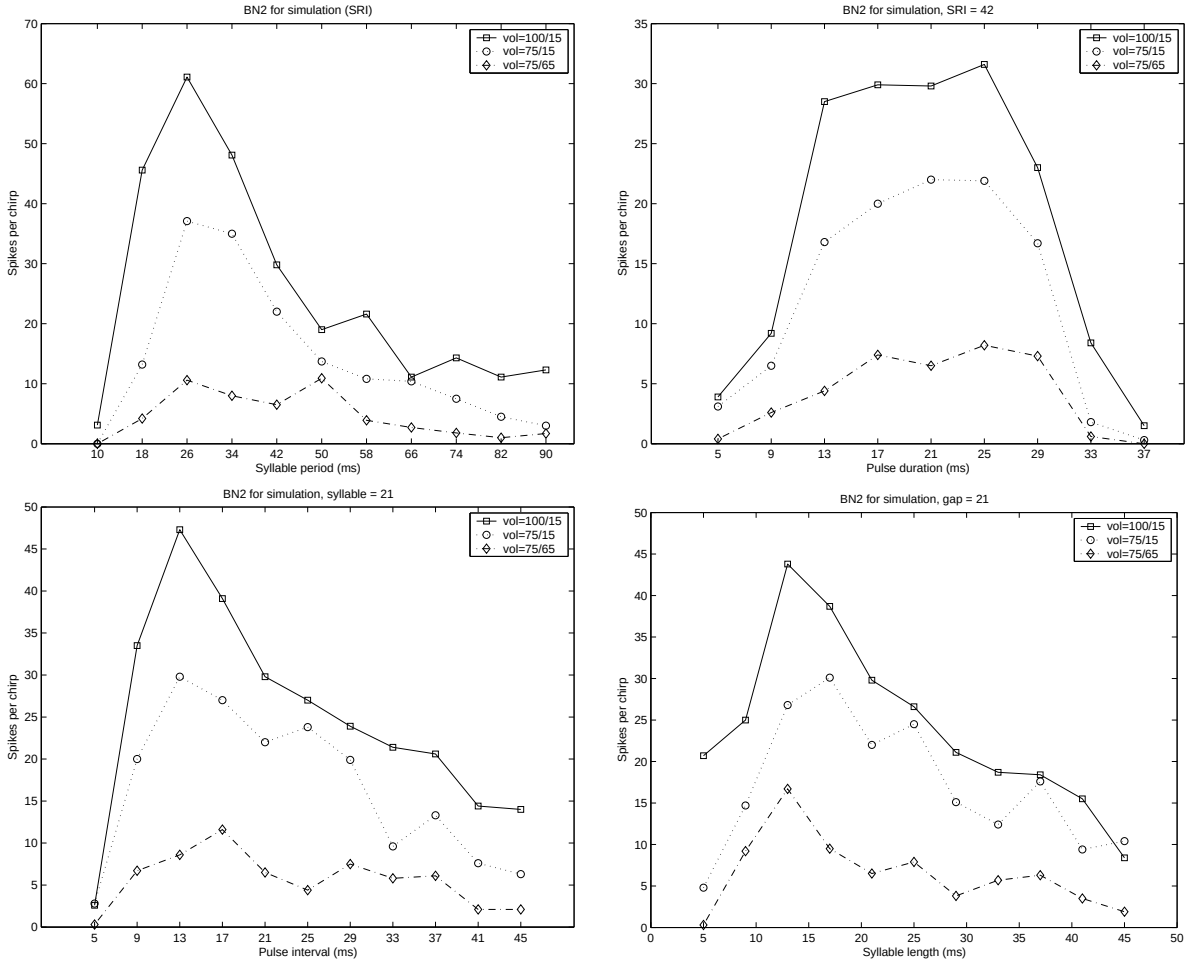

Figure 4. Model 1 for recognition. Spike rate output after filtering via BN1 and BN2. Top left, changing the syllable repetition interval ($pulse = gap$); top right, changing the duty cycle $pulse + gap = 42ms$; bottom left, changing the gap length; bottom right, changing the pulse length. This provides reasonable fit for the cricket data in figure 6.

6. Changing the observations

It is clear from the previous example that there is often, and importantly, an interaction between experimenting on the model and experimenting on the system: a good model will suggest new experiments to carry out on the biological system. It should be kept in mind, if simulation results fail to match biological data, that there might be some uncertainty about that data. It is important for the modeller to be as familiar as possible with the original reports and the methodology employed in the experiments. For example, (Schank et al. 2004) report how a pattern of behaviour noticed in their robot model led them to discover a previously overlooked pattern in the behaviour of rat pups. It is also important for the modeller to keep up to date with any new data that may be emerging. An recent exciting example for rat place cell models is the new data

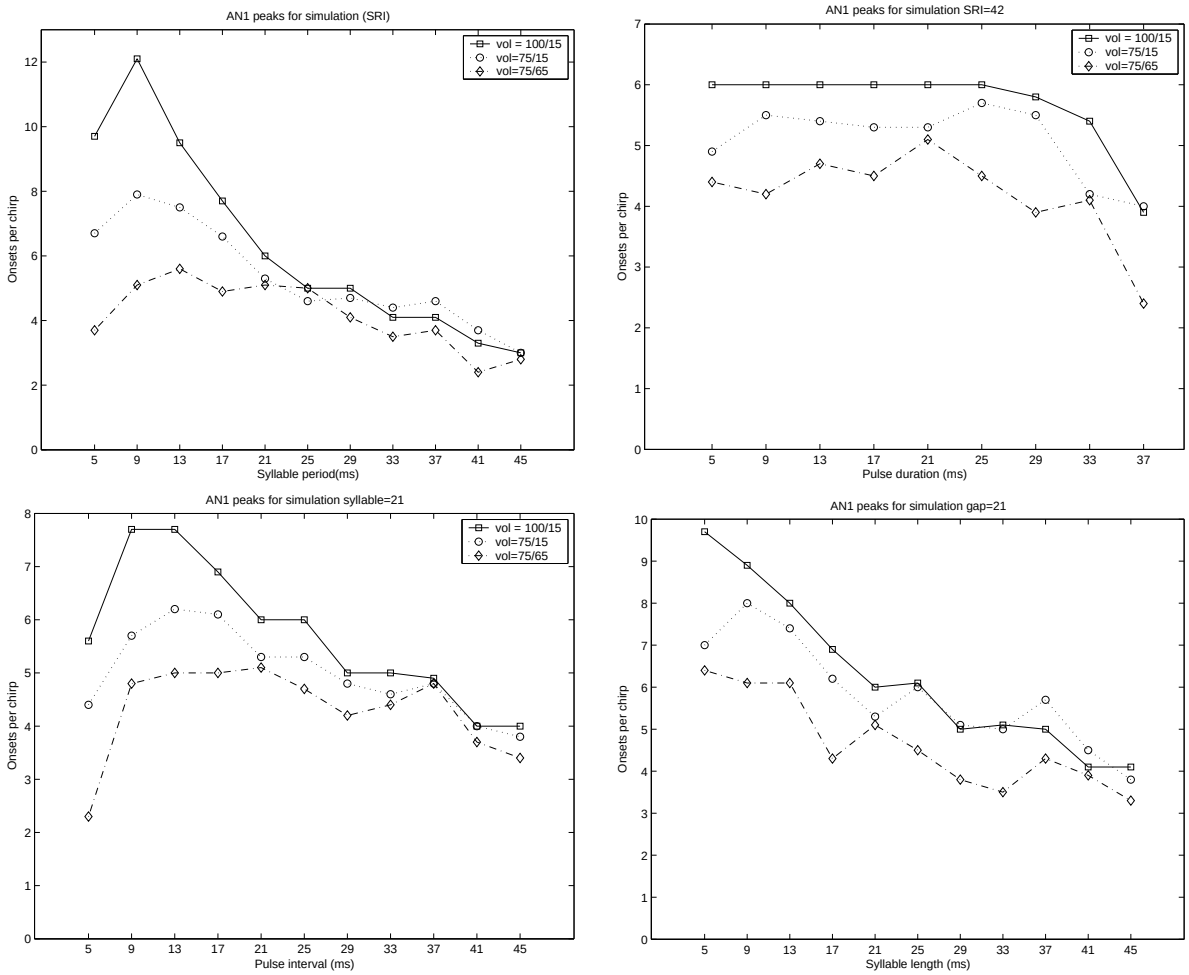

Figure 5. Model 2 for recognition. The number of 'peaks' in the firing rate of the AN1 or ON1 neuron detected in a fixed interval. Top left changing the syllable repetition interval ($pulse = gap$); top right, changing the duty cycle $pulse + gap = 42ms$; bottom left, changing the gap length; bottom right, changing the pulse length. The results for changing pulse duration and interval do not match the cricket data in figure 6.

revealing there is also a 'grid cell' system (Hafting et al. 2005). Although the strongest test of a model comes when it clearly predicts the result of some new experiment on the biological system, it is probably more common for the modelling process to simply reveal a gap in the data.

One lack of data that became particularly evident in previous robot modelling of the cricket was the poor characterisation of the dynamics of the response: e.g. how long is sound integrated before a turn is produced, how large a turn is made, and how often? Significant results have emerged in recent cricket experiments that use a much higher time resolution to analyse the behaviour (Hedwig & Poulet 2005). It was observed that there is a small but significant steering response to every pulse of sound within the song, within a short time interval (around 50ms). This strongly suggests that there is a turning

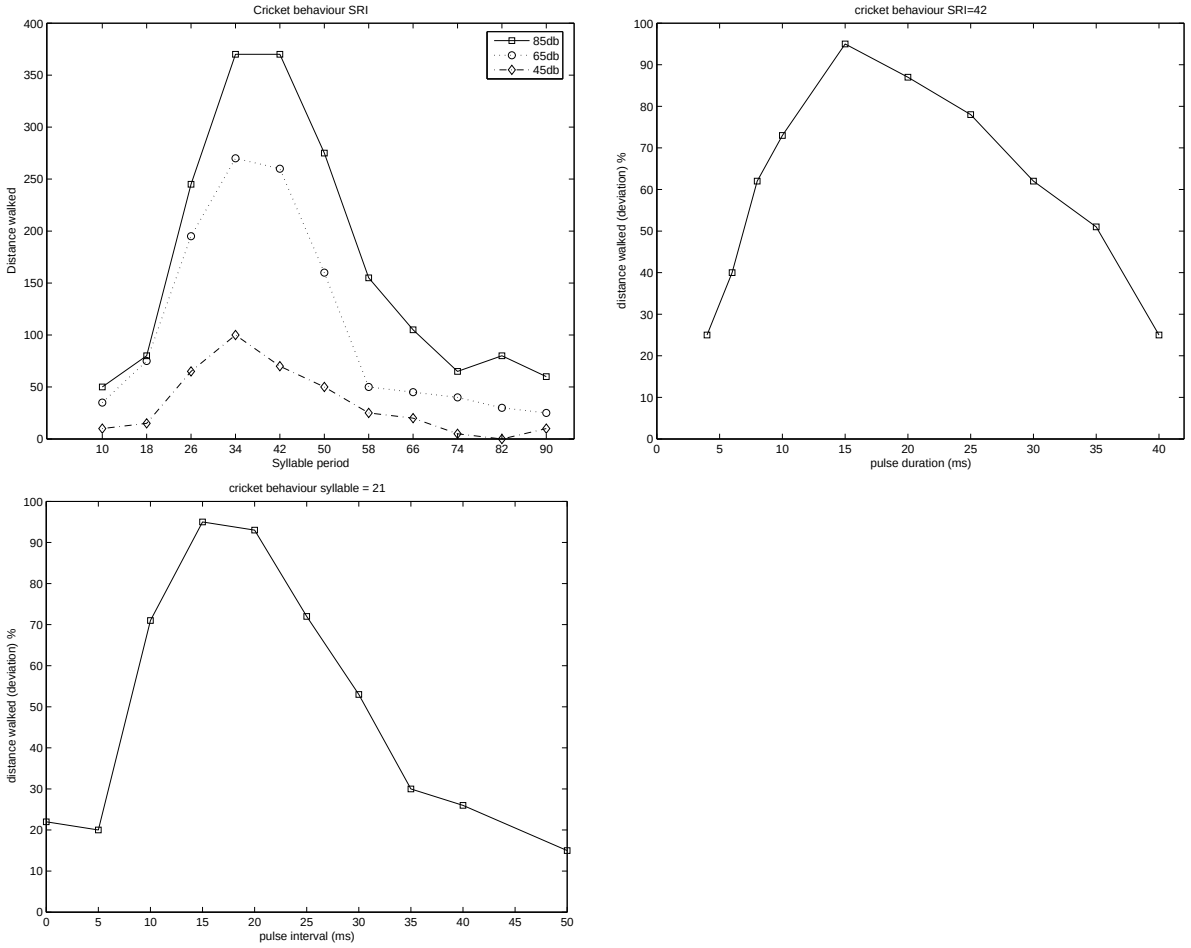

Figure 6. Cricket behaviour: top left changing the syllable repetition interval ($pulse = gap$); top right, changing the duty cycle $pulse + gap = 42ms$; bottom left, changing the gap length; the results for changing the pulse length are not known. Data from Hedwig and Poulet, personal communication.

reflex which does not include a stage for filtering the temporal pattern in the song, as the pattern (the syllable rate) cannot be detected from the first sound burst alone. Rather, it appears that the turning reflex is modulated, over a much longer time scale (2-5 seconds), by an independent process that filters for the temporal pattern (Poulet & Hedwig 2005). It was not possible to accommodate this data in the existing hypothesis implemented on the cricket robot; by imposing these experimental conditions we have found a failure to match, that cannot be compensated by changing the interpretation of the model output or the level, detail or accuracy of the representation. Rather, it requires a change of hypothesis.

The issue can be summarised as "Know your data!":

- Results frequently get simplified in the retelling. Don't rely on reviews or textbooks, but read original experimental reports. Be aware of how the data was obtained and what limitations there may be in the intepretation, and what exceptions exist.

- Always look out especially for data that will contradict your model. It is much easier to notice the confirmatory evidence.
- Keep up to date. If possible, work closely with the person doing the experiments. Consider whether you can do any of the experiments yourself.

7. Changing the hypothesis

Our most recent experiments on robot phonotaxis incorporate a change to the basic hypothesis about the mechanisms in the cricket (figure 7), based on the data discussed above. We suggest that there is a relatively direct, or 'fast' connection from the thoracic AN1 neurons to the motor control of turning. This connection is modulated by disinhibition via the brain neurons BN1 and BN2 which filter for the temporal pattern in the sound in the same way as before. Some interesting issues have been raised from initial tests on this model. It is difficult to obtain similar dynamics for the onset and offset of the response by varying the synaptic parameters within plausible ranges, suggesting that the modulation mechanism may require a different mechanism such as neuromodulator release. It is not clear from the data so far whether there is likely to be bilateral recognition, as in the current model, or whether the input from the two sides is combined in a single recognition process. I.e. a number of predictions or issues for further experimentation on the cricket have been raised.

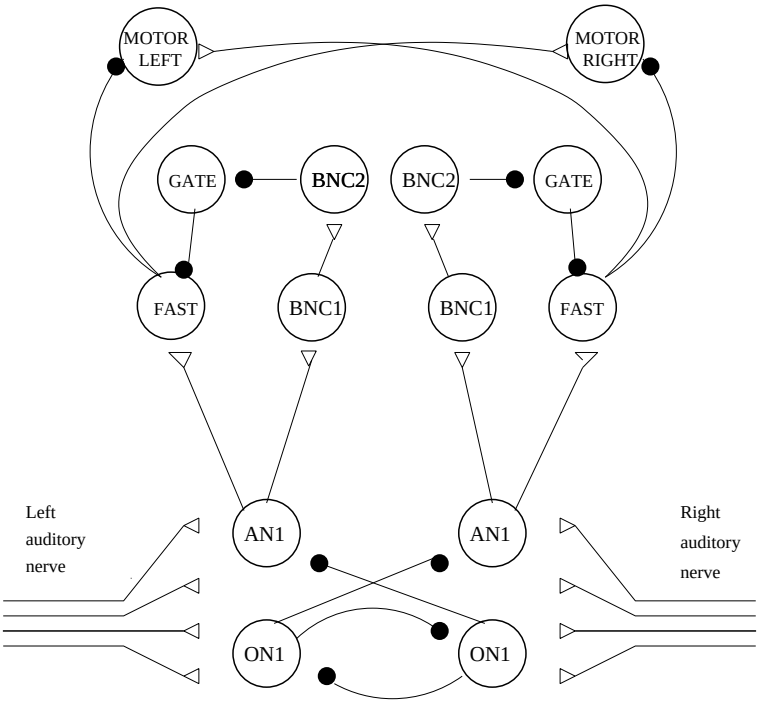

Figure 7. An alternative hypothesis for cricket phonotaxis

Note that this revision of the hypothesis does not imply that we need to restart the modelling process from scratch. We can re-use nearly all the infrastructure of the

implementation and analysis methods, and some results from the previous model are still relevant e.g. the mechanism for recognition has not changed. This emphasises the fact that the critical hypothesis itself only forms a small part of what is built in a biorobotic model, and that many other factors contribute to the output behaviour observed.

The strategy of holding the implementation, test conditions and intepretation constant, and changing the hypothesis, can be very useful even when there is not the clear motivation of a new or alternative explanation to be tested. Rather, one can carry out 'control' experiments in which some supposedly critical element of the hypothesis is ommitted, and check that the output changes in the expected way. In this way it can be possible to separate the effects of the implementation and the hypothesis on the outcome.

In summary, modelling is nearly always a process of going around the loop multiple times, as the implementation process raises questions that require further data, or suggests modifications are needed to the hypothesis; or the comparison of simulation and target behaviour leads to new hypotheses, and so on. Finding that your hypothesis is incorrect should be seen as an opportunity for further discovery.

8. Conclusions

In this paper I have outlined some of the complexities involved in validating biorobotic models. Most of the points apply to any kind of modelling, and indeed to problems in using experiments to test hypotheses. The essential point is to be aware of how the apparatus of modelling contributes substantially to the output and to the comparison with target behaviour, and thus needs to be kept in mind when attempting to draw conclusions from this outcome about the support or falsification of the hypothesis.

Thus a useful exercise when evaluating any account of modelling is to pose the corresponding questions:

- How valid is the representation? On what assumptions is it based? To what extent are the subcomponents also shown to be valid? Might a different implementation result in different behaviour?
- How do the experimental conditions used to test the model and the target system compare? What boundary condition assumptions are made in setting these up, and what might happen if they were exceeded?
- What is the basis for claiming equivalence between the simulation output and the target behaviour? How arbitrary is the labelling? What mapping factors are at the disposal of the modeller to improve the match?
- Can the experiments on the simulation actually be compared to data on the system, or are equivalent experiments prohibitively difficult, and if so, what can be learnt?
- How reliable is the experimental data? Are there relevant new results?

- How closely can the behaviours be compared, and to what extent is this simply dependent on parameter tuning in the model? What range of data can be accounted for, and have predictions been made and confirmed or falsified?

Moreover, all these issues can enter into the decision on how to choose a target for modelling in the first place. That is, it is useful to ask, what kind of representation will be possible and what assumptions will have to be made to build it? Are complementary representations possible? Can the same experimental methods and conditions be applied to the model as to the system itself? What is the quality of the available biological data, both for more or less detailed modelling, and for comparison with the model output? What alternative hypotheses might be compared using the same model infrastructure, and will predictions for the biological system be produced?

Finally, although this discussion has considered the modelling process as separate steps, it should of course be recognised that the separation between theorising, implementing, and experimenting is not always clear-cut. Changing the representation frequently involves some change to the hypothesis, as aspects require refining, simplifying or elaborating to make the implementation possible and functional. Tuning of model parameters, through experimentation and comparison to data, is a process in which free variables of the hypothesis may become fixed, or in other words, the hypothesis made more precise. Decisions about the test conditions may reflect changes in the intended scope or generality of the proposed mechanism. Nevertheless, keeping in mind the different aspects of the process as outlined in this paper should help to avoid some of the potential pitfalls.

Acknowledgments

Thanks to Guglielmo Tamburrini and Edoardo Datteri for discussions contributing to this paper, to the anonymous reviewers for their useful commments, and to Berthold Hedwig and James Poulet for the cricket behavioural results. Work described in this paper has been supported by grants from BBSRC, EPSRC, and the EC IST Sixth Framework Programme

References

Arleo A & Gerstner W 2000 *Biological Cybernetics* **83**, 287–299.
Braitenberg V 1984 *Vehicles: experiments in synthetic psychology* MIT Press Cambridge, MA.
Burgess N, Donnett J, Jeffery K & O'Keefe J 1997 *Phil. Trans. Roy. Soc., London B* **352**, 1535–1543.
Cartwright C 1983 *How the Laws of Physics Lie* Oxford University Press Oxford.
Grasso F, Consi T, Mountain D & Atema J 2000 *Robotics and Autonomous Systems* **30**(1-2), 115–131.
Hafner V V 2005 *Adaptive Behavior* **13**, 87–96.
Hafting T, Fyhn M, Molden S, Moser M B & Moser E I 2005 *Nature* **436**, 801–806.
Harding S 1976 *Can Theories be Refuted* Reidel.
Harman G 1965 *Philosophical Review* **74**, 88–95.
Hedwig B & Poulet J 2005 *Journal of Experimental Biology* **208**, 915–927.
Horchler A, Reeve R, Webb B & Quinn R 2004 *Advanced Robotics* **18**(8), 801–816.

Horseman G & Huber F 1994 *Journal of Comparative Physiology A* **175**, 389–398.
Ijspeert A, Crespi A & Cabelguen J 2005 *Neuroinformatics* **3**, 171–196.
Kanzaki R, Nagasawa S & Shimoyama I 2005 *Chem. Senses* **30**(suppl 1), 285–286.
Koch C 1999 *Biophysics of Computation* Oxford University Press, Oxford.
Lambrinos D, Mller R, Labhart T, Pfeifer R & Wehner R 2000 *Robotics and Autonomous Systems* **30**, 39–64.
Laudan L 1998 *in* M Curd & J Cover, eds, 'Philosophy of Science: the central issues' W.W.Norton & Company New York.
Lipton P 2004 *Inference to the Best Explanation* Routledge London.
Mataric M 1990 *in* J.-A Meyer & S Wilson, eds, 'Proceedings, From Animals to Animats: First International Conference on Simulation of Adaptive Behavior (SAB-90)' MIT Press Cambridge pp. 169–175.
Nabatiyan A, J.F.A. P, de Polavieja G & Hedwig B 2003 *Journal of Neurophysiology* **90**, 2484–2493.
Popper K R 1968 *Logic of Scientific Discovery* Hutchinson London.
Poulet J & Hedwig B 2005 *PNAS* **102**, 15665–15669.
Reeve R & Webb B 2003 *Philosophical Transactions of the Royal Society A* **361**, 2245–2266.
Schank J C, May C J, Tran J T & Joshi S S 2004 *Adaptive Behavior* **12**(3-4), 161–173.
Schildberger K 1984 *Journal of Comparative Physiology* **155**, 171–185.
Tamburrini G & Datteri E 2005 *Minds and Machines* **15**, 335–358.
Webb B 1995 *Robotics and Autonomous Systems* **16(2-4)**, 117–134.
Webb B 2001 *Behavioural and Brain Sciences* **24**(6), 1033–1094.
Webb B & Scutt T 2000 *Biological Cybernetics* **82(3)**, 247–269.