DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 40 (53) 2015

DOI: 10.1515/slgr-2015-0003

Marcin Miłkowski
Institute of Philosophy and Sociology, Polish Academy of Sciences

EVALUATING ARTIFICIAL MODELS OF COGNITION

Abstract. Artificial models of cognition serve different purposes, and their use determines the way they should be evaluated. There are also models that do not represent any particular biological agents, and there is controversy as to how they should be assessed. At the same time, modelers do evaluate such models as better or worse. There is also a widespread tendency to call for publicly available standards of replicability and benchmarking for such models. In this paper, I argue that proper evaluation of models does not depend on whether they target real biological agents or not; instead, the standards of evaluation depend on the use of models rather than on the reality of their targets. I discuss how models are validated depending on their use and argue that all-encompassing benchmarks for models may be well beyond reach.

Keywords: modeling, mechanism, mechanistic models, scaffolding, explanatory focus, idealization.

1. Introduction

In the modeling ecosystem of cognitive science, there are various kinds of models. Some are used merely for exploration, some serve explanatory purposes, while others provide a better understanding of cognitive phenomena. In this paper, I focus on the evaluation of artificial models of cognition. Artificial models are to be contrasted here with model organisms that can serve as proxies for studying cognitive phenomena. For example, in behaviorism, rats and pigeons were popular model animals; in today's neuroscience, macaques have become important, not to mention, obviously, human beings, who may be studied as proxies for all cognitive systems. In other words, artificial models are simply non-biological models, and they can take the form of computer programs, robots, trained neural networks, and so forth.

One field where artificial cognitive systems are studied is Artificial Intelligence. Another is cognitive robotics. There are currently two strands of

ISBN 978-83-7431-440-4 ISSN 0860-150X

research in robotics relevant here: on the one hand, there is research on *animats*, or possible creatures, which is supposed to provide insight into the principles of cognition or behavior (which has roots at least as deep as Tolman's 1939 "schematic sowbug"); on the other, there is robotic simulation of animals intended as explanation of real biological systems. The claim that animats are genuinely explanatory of biological systems is controversial because animat models do not correspond directly to existing biological agents (see Webb, 2009). So should one dismiss all animat research as producing mere gimmicks?

This controversy has deep roots in cognitive research. In the 1950s and 1960s, investigators pursued several explanatory strategies: one was cognitive simulation, which proceeded via building complete computational models such as the General Problem Solver (Newell, Shaw & Simon, 1960); another was building artificial neural devices such as the Perceptron (Rosenblatt, 1958) or robotic animals (Walter, 1950); and still another was what was to become cognitive psychology, called "information-processing psychology" at the time (for exemplars, see Miller, 1956; Broadbent, 1958; or Sperling, 1960). Information-processing psychology did not attempt to build complete implementations at all, but investigated the structure of psychological processes as based on information-processing considerations, and some psychologists, otherwise supportive of information-processing accounts of cognition, were quite critical of artificial modeling (Neisser, 1963). As cognitive psychologists used the kind of experimental evidence that was methodologically strict and widely accepted in psychology, information-processing psychology paved the way for recognizing cognitive research in general as valuable.

As there was no consensus about the value of artificial models in the 1960s, there is none today: animat modelers obviously do not agree with Webb (Beer & Williams, 2009). They usually defend the value of animat research by arguing that animats are not relevant to theories of cognition in the same way as robotic simulations are (Barandiaran & Chemero, 2009), and that different standards should be used for evaluating them. This leads naturally to what is my focus in this paper: What are normatively and descriptively adequate ways of evaluating artificial models of cognition?

Before I go on, some general remarks are in order. Artificial cognitive systems may take the form of a computer simulation or a physical entity. This distinction does not, however, correspond exactly to two kinds of simulation that are traditionally distinguished, namely representational and immediate ones (Krohs, 2008). Representational simulations are complex representations of a target phenomenon. For example, a digital simulation

of the weather in a computer is representational. Note that only a finite number of features are represented: no simulation can represent all the physical features of rain, so some of its features cannot be found in the simulation. Immediate simulations are used to model the target directly using physical resources; for example, a wind tunnel in aerodynamic research actually has aerodynamic properties, and a robotic model of an imaginary animal actually moves in space. But, at least according to some researchers, a computer simulation that instantiates an artificial cognitive system may just as well be classified as immediate. The researchers in question think that computer simulations of cognition are like a jazz improvisation produced by a computer: as information, it is indiscernible from a real improvisation (Dennett, 1981).

Here, I put aside the question of whether a cognitive computer simulation really is cognitive or just represents a cognitive process. What I am interested in is how one should treat artificial models of cognition. Is more detail better? Or maybe models should be more general? Are some models completely divorced from reality, as Webb seems to suggest? Or maybe they are, contrary to appearances, representations of some observable cognitive phenomena? Below, I will argue that these questions are answered differently depending on the use one makes of a model. Some artificial models cannot have explanatory uses, and in that case, they cannot be evaluated based on their representational power only.

The structure of the paper is as follows. In section 2, I will discuss important uses of artificial models of cognition. The list I provide is not supposed to be exhaustive, though it seems to lend some empirical support to the claim that evaluation standards depend on the use of models. I point out that the relationship between models and theories is more complex than usually presupposed, which makes it more difficult to assess the usefulness of modeling efforts. In section 3, I claim that universal benchmarks for models are out of the question and that they can actually hurt the progress in research on cognition. At the same time, I will argue that explanatory uses of models – be they biologically realistic or not – are possible only when models are representational of their targets. In this respect, there is no difference between robotic models of real animals and animats. Nevertheless, there is some truth to Webb's dictum that animats may have precious little to tell us about real animals: even if they are able to represent some features of animals, they may fail to be genuinely informative. To justify this claim, I will introduce a distinction between the model's intended focus and its scaffolding. The distinction will be used to justify the point that only some parts of models are crucial in their evaluation; I will also introduce

non-descriptive models, in which the distinction may be drawn but whose evaluation is different. In conclusion, I claim that modeling in cognitive science does not rely on a single standard of evaluation; instead, the standards employed depend on the use made of the model by the modeler. In other words, evaluation is relative both to the goals of the modeler and to standards in the scientific community, and there are no general, use-independent benchmarks for all cognitive models.

2. Uses of Artificial Models

Just like standard idealized models in science, models have various applications in scientific investigation. In this section, I discuss some (but surely not all) uses of artificial models in cognitive science. They range from prediction and explanation, to formalization of verbal theories, to conceptual exploration and thought experiments, to engineering purposes, especially in AI. I will claim that, depending on one's purpose, one and the same artificial model can be assessed differently.

2.1. Models as predictive and explanatory tools

The primary use of modeling is to describe, predict, and explain various phenomena. Usually, models are said to have those functions as long as they correspond to the phenomena in question; however, according to some authors, models may sometimes be treated as fiction without being explanatorily irrelevant (Suárez, 2009; Godfrey-Smith, 2008). It is notable that the latter position is largely developed in connection with non-explanatory uses of models (such as conceptual exploration), and there is an obvious objection against fictionalism about explanatory models: if they are fictions, how are they supposed to explain the non-fictional (Levy, 2012)? Due to lack of space, I will not go into the details of this important discussion here. For my purposes the most important thing is that even fictionalists think that models are representations, i.e., they refer (or fail to refer) to target systems. But although, in the case of many modeling efforts, the target systems do not exist at all, we are able to say what properties they would have if they did exist. In other words, the only difference between fictionalists and realists is that realists usually take models to be idealized, truth-constrained representations of reality, while fictionalists call idealized models "fictions". But they share the same assumption, namely that there is something about models that allows them to represent reality, though fictionalists stress that the representational power of models cannot be captured merely by isomor-

phism or structural resemblance (Suárez, 2003). Indeed, this is what realists also claim: the intention of the modeler and the practice of the research community count as well (Weisberg, 2013). However, there has to be a certain structural relationship between the model and the target system; otherwise, these would not be models but symbols, viz. representations whose semantic values depend merely on convention.

Let me turn to artificial models of cognition. Most commonly, they are computational models, expressed in terms of programs and input data, trained artificial neural networks, or other computational tools. There are two ways in which formal computational models of cognition are usually taken to correspond to cognitive phenomena. First, they may be *weakly equivalent* to a cognitive process, in that they only describe its input and output data. Second, they may be *strongly equivalent*, in which case they also correspond to the process that generates the output data. These notions have been used in the methodology of computer simulation since the 1960s (Fodor, 1968, chapter 4). Similar terminology has been introduced by Bernard Zeigler in his classic theory of modeling and simulation (1976): a model is said to be *replicatively* valid if it can generate the output data from known input data; it is *predictively* valid when its output corresponds to yet unobserved data, and *structurally* valid when the structure of the model corresponds to the operations of the real system being modeled. Zeigler's predictive validity is equivalent to Fodor's weak equivalence, and structural validity to strong equivalence. (Note that Weisberg, 2013, p. 41, makes the same distinction but uses the somewhat confusing terminology of the *dynamical* and *representational* fidelity of models.)

These distinctions correspond neatly to the possible uses of models. Replicatively valid models can only describe the experimental evidence gathered so far; predictively valid models are also able to extrapolate to new input and output data. Both, however, are too weak to explain the target systems fully: the structure of the model that generates the output data need not correspond to the real structure of the target system. Hence, we need structurally valid models to be able to explain the behavior of systems with computational modeling.

The latter claim can be disputed by proponents of the functionalist account of computational explanation (Cummins, 1983), who think that a model is explanatory as long as it suffices to generate new output data based on new input data, i.e., as long as it is predictively valid. While the validity of these models is not always coincidental and may stem from the fact that they describe the explanandum phenomenon in sufficient detail, it remains controversial that they explain the phenomenon at all (see Piccinini

& Craver, 2011; Miłkowski, 2013). The basic argument used to support the claim that only structurally valid models of cognition are explanatory of cognitive processes is that merely predictively valid models may produce the same output as a cognitive process, but in a completely different way. By analogy, although both a bird and a jet can fly, a jet would not be a good explanatory model of how birds can fly, because the mechanism producing flight in a jet is not sufficiently similar to the mechanism producing flight in a bird.

This brings me to an important distinction to be made, namely between the model's *intended focus* and its *scaffolding*. The intended focus is what the modeler intends to stand in correspondence to the model's target system. The scaffolding plays a merely supportive role, and is tacitly ignored during the model's validation. For example, a recent massive model of the brain, SPAUN (Eliasmith et al., 2012) is able to perform eight diverse cognitive tasks. SPAUN has been implemented in Nengo, which is a simulation framework written mostly in Java programming language. But the fact that Nengo is executed by a Java Virtual Machine (VM) is irrelevant for its validation as a biologically plausible model of the brain. Obviously, nobody expects anything strictly analogical to a Java VM in the brain. It's only a certain pattern in SPAUN's information-processing architecture that is supposed to correspond with the brain. Hence, the Java VM is just scaffolding: a necessary component of the model, as Nengo cannot run without it, which is ignored during the validation of the model.

In weakly equivalent models, the structure of the computational process as implemented by the model is treated as scaffolding. The intended focus is just the input and output data. In the case of structurally valid, strongly equivalent models, the intended focus of the model includes the structure of the computational process as well.

What the modeler treats as the scaffolding and the intended focus is essential in assessing the value of the model. Again, if we were to think that SPAUN's implementation in Java is part of the intended focus, the model would be biologically implausible. It is plausible only because the intended focus does not include the Java implementation. Note that if it turns out that what was treated as scaffolding is also in correspondence with the explanandum phenomenon (for example, it turns out that, contrary to appearances, some brains actually do contain Java VMs), then the modeler can change his or her interpretation of the model.

Unfortunately, it is sometimes difficult to tell the intended focus and scaffolding apart. The interpretation of models relies on implicit assumptions in the practice of researchers and is rarely brought to the fore. What

is more important, modelers usually need to add multiple new assumptions to the specification they had at the beginning of the modeling process just to implement a model, and such ad hoc additions are difficult to identify when observing the behavior of a model (Lewandowsky, 1993). The ad hoc additions usually belong to the scaffolding of the model, but they may actually correct the mistakes made in what is thought to be the intended focus.

Think of a simplistic model of human reasoning. It is a well known fact that people consider *modus tollens* slightly less plausible than *modus ponens* as a rule of deduction (Oaksford, Chater & Larkin, 2000). One could easily implement a model that replicates that experimental data simply by having classical reasoning rules and a correction rule that would remove some of the *modus tollens* results (for example, by removing 20% of cases to replicate the experimental data). This is an ad hoc model, but it could be predictively accurate. Of course, the ad hoc rule is just made up to fit the experimental data, but one could argue that how the model has been built is irrelevant to its validity, as the context of discovery is different from the context of justification, and that the structure of the model makes it actually explanatory of human reasoning. But just because the ad hoc addition does not correspond, as far as I know, to what happens during everyday reasoning processes, the model will not be structurally valid if the ad hoc addition is taken to be part of the intended focus of the model. If, however, it is declared to be part of the scaffolding, then one should ignore the results of correction in the data. This is easier said than done; of course, it might be technically possible to analyze my toy model (if it's not hopelessly obfuscated by an evil scientist), but a very complex production system may easily contain thousands of rules, some of which are merely supportive and not psychologically plausible. But just because of this fact, very complex models with ad hoc or special supportive additions cannot safely be said to be structurally valid. It may be probable that they are, but to say so, we need to analyze the models' behavior in detail instead of simply taking the results of their operation at face value.

2.2. Models as unambiguous formulations of a theory

Models have obviously other uses than description, prediction, or explanation. One prominent advantage of modeling that most proponents of modeling in cognitive science and psychology cite is that models "serve as unambiguous formulations of a theory" (Frijda, 1967). Merely verbal theories may lack clarity or contain serious gaps that go unnoticed just because certain questions were not asked. For this reason, it seems reasonable to

claim that computational – or robotic – implementations require that an extant verbal theory become more detailed. As convincing as this claim may sound, there are some difficulties in justifying it. In addition, some theories of smaller scope, for example so-called microtheories of experimental psychology, might be equally unambiguous as computational models and they can be expressed mathematically. In other words, computational artificial models of cognition are not the only means to achieve unambiguous formulations of theories.

But before one can talk of models and their relationships to theories, it has to be made clear what is meant by the terms "theory" and "model" in the first place. Admittedly, there is no widespread agreement as to what counts as a theory in cognitive science; but it seems that most modelers use the notion to denote some scientific representation that is more general than a model, while the latter describes a single entity or phenomenon. I will use this rough distinction in what follows, even if the terms remain somewhat vague and there are problematic cases in between (for example, microtheories proposed by Newell & Simon, 1972, are at the same time computational models in the form of computer programs).

The most important problem with the claim that models are unambiguous formulations of a theory is that the relationship between a theory and a model is definitely not of a logical deduction; so a model is not a *formulation* of a theory at all. Also, even if it were logically deducible from a theory, it would have a smaller scope, so it would be its *instantiation*. There are two ways one could deal with this difficulty. One could simply say that models are kinds of theories, a move that is recommended by Sun (2009). But, in this case, no previous theory gets more unambiguous during modeling; it is just that modelers produce a new piece of research. This research obviously does not come from nowhere; so the new theory would be a stricter replacement for the older one. In this sense, models (new theories) could be (somewhat clumsily) said to be stricter formulations of old theories, but what is rather meant is that new theories are more precise than previous theories.

Alternatively, one can interpret Frijda as saying that theories themselves become more precise during the development of the model. This is, however, true only if the model's accuracy is achieved not ad hoc. In other words, we can say that models can make the theories more precise only if the intended focus of the model includes some additional information that disambiguates the previously existing theory. If it's just the scaffolding that helps make the model clearer than the theory, then we should ignore the scaffolding when assessing the model.

2.3. Models as means to check the consistency and completeness of theories

Yet another use of models is to check the consistency and completeness of theories by implementing them (Frijda, 1967; Farell & Lewandowsky, 2010). Again, we can understand the relationship between models and theories at least in two ways, just like in section 2.2. Also, owing to multiple ad hoc additions, it may be difficult to say if a discovered inconsistency is part of an older theory or just a side-effect of model-building. For this reason, in this use, the distinction between the intended focus and the scaffolding remains equally important as before.

One thing needs to be noted here. The widespread belief (endorsed for example by Frijda) that it is impossible to implement a contradictory theory is simply false; while programming languages disallow syntactic errors, semantic inconsistency is possible: It is certainly one source of bugs in the code of programs. So the admissibility of using a computational model of a cognitive process as a consistency check depends on multiple factors, and in general, for very complex models, it may not be feasible to perform a full proof of their correctness. For this reason, models as consistency checkers can, at best, be used as fallible heuristics. There might be a problem with a theory when it cannot be implemented in a computational model; but it may also be the fact that the modeler lacks the imagination and skills required to develop the model. Worse still, a model may be so complex that it may contain mistakes inherited from the theory, but they simply won't show up during the standard validation vis à vis empirical evidence.

2.4. Models as guides in discovery and search for other models

Models are also used to guide the discovery and search for models (Frijda, 1967; Barandarian & Chemero, 2009). Namely, by running computer simulations and reconfiguring robots, one can perform thought-experiments with systems that are too complex for people to understand without any external aids. For this reason, Di Paolo, Noble, and Bullock (2000) call these thought-experiments "opaque"; Dennett (1991) recommended using artificial life exactly for the same reason. By running experiments on artificial systems, modelers may discover that their initial intuitions were actually wrong. These might be not only intuitions but complete theories as well: they may uncover unexpected implications of a theory. It might be illuminating even when no experimental evidence was taken into account, and may lead to interesting new hypotheses (Frijda, 1967).

This use of models is especially stressed by Barandiaran and Chemero (2009) in their vindication of animat modeling. It is understood quite

broadly to encompass even such applications as comparisons between "explanatory paradigms" and exploration of potential interactions between theories. The talk of comparing paradigms might seem slightly surprising, given that paradigms for Kuhn (1970) were understood as incommensurable, therefore incomparable. The notion of "paradigm" is apparently used here in a weakened sense to mean a general methodological approach. What Barandiaran and Chemero probably mean by "comparing" is building comparable models to instantiate assumptions of different paradigms in this weak sense – for example, the symbolic paradigm and the connectionist paradigm in cognitive science. It is possible, for example, to build LEGO robots driven by these two modeling methodologies and compare their performance. But such comparisons, again, rely on a somewhat problematic distinction between a theory and a model, and this makes the results of such comparisons somewhat fallible.

Another point worth stressing as regards models used to guide discovery is what Braitenberg (1984) called the law of the uphill analysis and downhill invention: namely, artificial models are (usually) easier to understand, so building them helps to analyze the real systems. Note that there is an important exception to Braitenberg's "law", called Bonini's paradox: sometimes modeling is more difficult to understand than real systems, which is especially true of some connectionist models (Dawson, 2004). But even in complex systems, changing the parameters of a model and running the experiments again might be illuminating; this is what Cleeremans and French (1996) call "probing the model": They accentuate the importance of changing the parameters of running models for understanding phenomena.

2.5. Uses of models whose intended targets do not exist

All the above mentioned uses are common to models with a real target system and ones without one. But if there is no observable target system, it may seem that artificial models of cognition are not explanatory: at least they cannot be explanatory of something that does not exist. After all, one cannot explain why Sherlock Holmes solved a murder mystery because Sherlock Holmes, alas, never existed. Barandiaran and Chemero (2009) defend the view that there might be what they call "generic models" that "stand in abstract and generic correspondence with multiple phenomena", in contrast to functional models, which correspond to behavior (or input/output functions) of targets, and mechanistic models, which require one-to-one correspondence with a target. Besides that, they also distinguish "conceptual models" that correspond only to a theory.

As I already mentioned in section 2.2, the relationship between theories and models is not of correspondence at all (and need not be), as theories have a different scope, grain, and level of generality. Hence, theories do not usually contain all kinds of detailed descriptions and the scaffolding needed for a complete model to work. But for the sake of argument, let us assume that there is a way to understand the notion of correspondence as used by Barandiaran and Chemero. Would that mean that models need to be checked or validated by testing whether they correspond (in some sense) to a theory? I doubt that.

Let me elaborate. Even if scientists build models that describe imaginary animals, they usually think that these models are not just literary fiction, created (only) for fun (which is what even fictionalists would accept). They seem to assume that there is a body of knowledge which is not reducible to formal or mathematical principles of the models they build; in short, these models are supposed to tell us something about observable phenomena, just like other idealizations in science. Some of the models, including conceptual models, while not meant as direct representations of any real observable target phenomena, instantiate at least some properties of observable phenomena in a useful way.

One interesting example is given by Barandiaran and Chemero: "The Baldwin effect (Baldwin, 1896) ... was nicely demonstrated by a computer model by Hinton and Nowlan (1987) and gave rise to a revival of the subject (Weber & Depew, 2003)." The computer model was not a realistic model of life at all, but shared enough properties with biological evolution to be useful. The target of the model was not an observable phenomenon; but representational properties of the model were such that they described a range of biological phenomena anyway. In other words, a single intended target system does not exist; but the intended *focus* does correspond to features of real systems. There are entities that have at least some properties described by the focus of the model, and this is why the Baldwin effect is worth studying in biology. Otherwise, it would be mere science fiction. To wit, some artificial models *can* be explanatory about empirical phenomena by instantiating general principles. They do not need to correspond to a complete animal at all, if the complete animal is not what is supposed to correspond to their intended focus.

What about models whose intended focus is *not* supposed to correspond to anything real? For example, one may be simply interested in investigating possible outcomes of some configurations of a model; these possibilities might not match any single phenomenon nor have any observable intended focus. Similarly, traditional "sufficiency" explanations in cognitive science,

i.e., explanations that show what would be sufficient means for a cognitive system to perform a given task, may be understood as explorations of the possible. (Whether *explaining* the merely possible is actually relevant to science is another matter.) These explanations of the possible are not necessarily bound with any observable intended focus.

There are also models that are used in engineering to develop new designs, for example in robotics. Surely such robots did not exist in the past and the reason for building a model is not to explain or describe anything but to create a new entity that has some capacity. This is a strictly engineering use of an artificial model. Most successful work in Artificial Intelligence aims at developing new, effective tools. For example, contemporary machine translation, though related to cognitive science, is usually performed in a way that guarantees that a translation engine simply does the job. The operation of the engine does not explain much about human translation, if anything at all (we already know the input/output function anyway, even if we do not know how it is computed). But it may excel at performing the task. Note that merely engineering uses of models do not require the modeler to distinguish between the scaffolding and the intended focus; actually, these models may be simply *tools* rather models of anything, to be exact. In other cases, when models are supposed to match human performance, for example, they can be understood as weakly equivalent models in the traditional sense.

The uses I enumerated in this section of the paper do not constitute a complete list; it is just a partial taxonomy that may contain overlapping categories. Nonetheless, I think that it may guide our thinking about evaluating models: the most important point is that there are different, sometimes dramatically so, applications of artificial modeling.

3. Evaluating Progress

Depending on the intended use of the model, the method of evaluation varies. Let me start with theoretically unproblematic engineering applications. They usually have well-defined capacities that the model is supposed to display, and sometimes automatic benchmarking is possible. For example, a statistical machine translation engine might be evaluated by human translators who assess the fluency and the adequacy of the translated text as compared to the source one. As human evaluators are costly, several automatic metrics were proposed as a rough evaluation of results before they are handed over to humans. Defining criteria of fluency and

adequacy in a formal manner, as it turns out, is not a trivial task in itself, and popular benchmarks depend on a metric of similarity between the reference human translation and the machine-generated one. For example, one of the most popular metrics, BLEU (Bilingual Evaluation Understudy), considers how many phrases (sequences of words, called *n-grams*), overlap in both texts. Even if BLEU is known to diverge somewhat from results of human evaluation (though the correlation of human judgments and BLEU scores is actually quite high), it is used as a cheap evaluation tool during the development of machine translation engines (for more detail and an introduction to evaluation of machine translation, see Koehn, 2010, chap. 8).

In other words, although there are outstanding problems with the evaluation of machine translation, it is clear in principle what should be compared and why; what we need is a good mathematical definition of the metric to perform the comparison. Some other factors might also be taken into account, such as speed, adaptability, ease of use, or size of the system. All these are related to the technological use of the model.

Evaluating models whose use is not confined to engineering is by far more complicated. In cognitive science, traditional descriptive computer simulations were assessed by looking at how much they correspond to observable behavior (including input-output functions) or to observable behavior and known facts about the organization of the system. In other words, depending on the intended level of equivalence – strong or weak – one can compare the model and the target system, and the results of comparison can usually be expressed quantitatively.

Note that artificial models with a clear intended target can be validated vis à vis empirical evidence. So, to use the Baldwin effect example, if there is a similar process in biological evolution, one may verify whether it has properties predicted by the model developed by Hinton and Nowlan. In the case of cognitive models without a real intended target system, observable evidence will be related to selected cognitive *capacities* or *principles*. There is nothing especially difficult about evaluating such models, though it is not useful to create specific benchmarks before an encompassing theory is built.

Models that are not intended as explanatory tools but rather as proofs of possibility are to be evaluated in a different way. A proof of possibility may overthrow a theory that declared something to be impossible. Take for example McClelland and Rumelhart's (1986) model of past tense acquisition in English: it demonstrated that it is possible to learn morphology without explicit rules. The model was at first also intended as representing

an observable target (though with important simplifications; for example, the network used to learn past tense used only verbs, not complete sentences, on its input, and that constitutes a Galilean idealization). It was soon criticized (Pinker & Prince, 1988), but the proof that it is logically possible to account for morphology without any explicit rules seems to be still generally accepted. One does not need to treat the original model as descriptive of the real acquisition process but as a generic model of such phenomena; in such a case, this model is fine, criticisms of Pinker and Price notwithstanding.

But there is a perceived danger in changing the intended use of a model. One may object that to change the use of a model is to lower the criteria of evaluation. If it's always possible to change one's mind about how a model is supposed to be used, then it becomes easier to get valuable models and more difficult to go completely wrong. This is, it seems, one of the motivations behind Webb's (2009) criticism: it is much harder to make a biologically faithful model of a cricket than to create a simple imaginary creature along the lines of Braitenberg (1984).

Generic models as such simply contain less information – this is what yields generality at the price of depth. So, choosing theoretically relevant phenomena as intended targets is important; otherwise, the models are too general to be interesting. In other words, it's not enough for the proponent of a generic model to show that his or her model has an intended focus that corresponds to real principles or capacities. What the model shows should be non-trivial. For example, numerous models showing that complex behavior may be the outcome of very simple constituent operations, combined using simple principles, are now considered too trivial to be interesting. They may be correct as far as it's true that the complexity they produce is not spurious; this kind of result is hardly novel.

In other words, more generality is not always better, just as trivial details do not matter. Not all models without a biological intended target system are divorced from reality, as Webb seems to suggest. These models instantiate properties that modelers consider important for understanding; for example, principles of cognition or social interaction. Yet animats usually work in a very simplistic way, and for this reason, building them is not as beneficial for neuroethology as biorobotics is: To build a model of a real ant, one needs to know a lot about ants, and perform experiments that had not been performed before. The important thing here is that many animats may have large scaffolding and a very small intended focus. A case in point is StickyBot: a robot resembling a gecko that is supposed to explain how geckos are able to walk on the ceiling (Santos,

Sangbae, Spenko, Parness & Cutkosky, 2007). The problem is that we already knew how they do this without the model: it is the structure of the gecko's feet and van der Waals force that allows the animal to stay on whatever surface it wants to be. In StickyBot, the intended focus of the model includes only the feet, and that's precious little. The rest of the robot does not explain anything else and is simply gimmickry (Sanz & Hernández, 2010). In my terminology, the StickyBot model is composed mostly of scaffolding that captures people's attention. And it does: it won Time Magazine's Best Invention of the year 2006. In the section Toys, to be exact.

In general, a model's representational value depends on the ratio between the scaffolding and the intended focus. In cases where the scaffolding becomes bigger than the intended focus, there should be a good justification for such a decision. Otherwise, the model might be a mere gimmick, just like StickyBot. The representational value of a model is important for explanatory, predictive, and replicative uses of modeling; for models as tools that make theories more precise or are used to check their consistency, the representational value of the model itself may sometimes be ignored during evaluation (but the theory being checked or made more precise should not be empirically unsound, otherwise, there is no point in using the models).

What about models of possible cognitive systems in cases when nothing observable matches the intended focus at all? The models of this flavor, which belong to the field dubbed "Android Epistemology" by Glymour (1987), can be evaluated either in the light of engineering criteria, which makes them theoretically rather unimportant; or in the light of theories of cognition that they would conform to. The theories in question need to contain operational criteria of success of modeling; i.e., they have to define cognitive capacities in such a manner that they can be tested. But herein lies the problem: different theories conceptualize cognition, and cognitive capacities or processes, in dramatically different ways. For this reason, such artificial models will be theory-dependent in a fairly strong way. Their validity will depend on the empirical validity of a particular theory. Even Android Epistemology needs to be related to observable cognitive systems to be scientifically relevant.

To this, someone might object, in a Platonist vein, that exploring the space of possible cognitive systems does not require having any theories of cognition. After all, are all animats really so strongly connected to theories of cognition? The problem with such a position is that we would need to know first what kind of space the model is exploring. How can we tell that it

is a model of cognition, and not, say, of merely reactive behavior? Without a theory, artificial modeling of possibilities is blind. And with theories, it may be validated.

An example is in order. One of the contemporary hypotheses in cognitive science is that "mind is life", which is sometimes understood more literally as a claim that cognition relies on metabolic processes of some kind. This is an obviously troublesome position for robotic modelers. Should they accept it, they would need to build a robotic model that relies on some form of metabolism. Interestingly, Montebelli et al. (2010) created a robot that generates energy from unrefined renewable biomass. The problem with this robot model is that we do not know whether it operates successfully to achieve any goals. As such, it may be called just a motor engine that uses biomass to move. It is simply utterly unclear *which* property of metabolic processes is important to cognition and in what way: the modelers do not say that. Even if metabolism might be brought to bear on the autonomy of an agent, even if it is a source of normativity for the agent, it is still not obvious that replacing a standard battery with a biological source of energy makes any difference unless we understand the role of metabolism in cognition.

Does this mean that "mind is life" has been reduced to absurdity? Certainly, its opponents would be tempted to say so. But the relationship between the biomass-driven robot and the theory is not one of a logical deduction. Even if we discredit the model, the theory could still be upheld. This means that evaluating models can only be a fallible guide to evaluating theories; but some theories cannot be fully evaluated otherwise, as empirical evidence is too scarce or the phenomena too complex.

Even if there were any simple methodology to create benchmarks to evaluate artificial computational models, there would still be no simple way to assess the progress. The relative merits of models, if they could be compared using the same scale and with some ordering relation, depend themselves on theoretical progress, the breadth and scope of theories and their empirical validation. Simple evaluation is but a bureaucrat's dream. Universal benchmarks for all uses of models simply cannot exist.

This is true for most theoretical uses of artificial models of cognition. For checking the consistency of theories using modeling, we need fairly complete theories to start with. Otherwise, the inconsistencies found might be caused by the scaffolding built for the model to work. In these cases, however, the validity of the model depends simply on its correspondence with a theory, and assessing it is still an art, not a methodology.

4. Conclusions

In this paper, I have reviewed some uses of artificial modeling in cognitive science, and related it to different ways of validating the models. I introduced a distinction between the parts of the model that are relevant to the model's correspondence to target systems – namely the intended focus of the model, and the parts that simply support the model to help represent its intended target systems, which I dubbed scaffolding. The standard way of validation of models of cognition consists in assessing the match between empirical phenomena and the intended focus, i.e., in checking whether the data generated in the model matches the observation. For models built for theoretical purposes, for example to compare, check consistency, or otherwise explore theories, their validity depends either on a correspondence between model and theory, or – when the models are supposed to further progress in the empirical research – on empirical validity of the theory that they are models of.

One particular difficulty here is that theories in cognitive science are not easy to evaluate empirically in the first place, and they are usually validated using models. This might raise a worry that there is a vicious circle involved. I think this circularity may be broken simply by creating multiple independent models of a theory; if they lead to similar results, then their individual independent lies will intersect: the intersection would be the truth. Briefly, multiple models idealization (Levins, 1966) seems to be the way out of this circularity.

Artificial models are an important part of the modeling ecosystem. By looking at how modeling works in practice, and how modelers distinguish it from building theories, one can gain important insights about the structure of what is understood as theory in the sciences that peruse artificial models. There is at least an apparent tension between the received view that holds that theories are to be understood propositionally, ideally as axiomatic systems, and the practice of theorizing in such disciplines. Insofar as philosophy of science is supposed to provide normative guidance, it is important to reflect upon the discrepancy between the traditional norm of axiomatic theory building and the messy practice. There has not been enough stress on the heuristic role of models in theory building, nor on the treatment of models as fallible heuristics used for further discoveries and general theoretical exploration. At the same time, there is a plethora of models that seem to focus on fairly trivial phenomena, such as the emergence of complexity out of simple components. Such stock models have usually precious little to explain, and they do not seem to be useful for any further discovery.

Nonetheless, these are the topics that all deserve separate treatment in their own right.

Acknowledgements
Work on this paper was financed by the National Science Centre OPUS Grant under the decision DEC-2011/03/B/HS1/04563. The author wishes to thank the anonymous referee for helpful comments on the previous version of the paper.

REFERENCES

Barandiaran, X. E., & Chemero, A. (2009). Animats in the modeling ecosystem. *Adaptive Behavior*, *17*(4), 287–292.

Beer, R. D., & Williams, P. L. (2009). Animals and animats: Why not both iguanas? *Adaptive Behavior*, *17*(4), 296–302.

Braitenberg, V. (1984). *Vehicles, experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Broadbent, D. E. (1958). *Perception and communication.* Oxford: Pergamon Press.

Cleeremans, A., & French, R. M. (1996). From chicken squawking to cognition: Levels of description and the computational approach in psychology. *Psychologica Belgica*, *36*, 1–28.

Craver, C. F. (2007). *Explaining the brain.* Oxford: Oxford University Press.

Cummins, R. (1983). *The nature of psychological explanation.* Cambridge, MA: MIT Press.

Dawson, M. R. W. (2004). *Minds and machines: Connectionism and psychological modeling.* Malden, MA: Blackwell.

Dennett, D. C. (1981). Reflections on D. R. Hofstadter's "The Turing Test: A coffeehouse conversation". In D. R. Hofstadter & D. C. Dennett (Eds.), *The mind's I* (pp. 69–95). New York: Bantam Books.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338*(6111), 1202–1205. doi:10.1126/science.1225266.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329–335.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology.* New York: Random House.

Frijda, N. H. (1967). Problems of computer simulation. *Behavioral Science*, *12*(1), 59–67.

Glennan, S. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(S3), S342–S353.

Glymour, C. (1987). Android epistemology and the frame problem: Comments on Dennett's "Cognitive wheels." In Z. W. Pylyshyn (Ed.), *The robot's dilemma: Frame problem in Artificial Intelligence* (pp. 65–75). Norwood: Ablex.

Godfrey-Smith, P. (2008). Models and fictions in science. *Philosophical Studies*, *143*(1), 101–116.

Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, *1*, 495–502.

Koehn, P. (2010). *Statistical machine translation*. Cambridge: Cambridge University Press.

Krohs, U. (2008). How digital computer simulations explain real-world processes. *International Studies in the Philosophy of Science*, *22*(3), 277–292.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, *4*(4) (July), 236–243.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431.

Levy, A. (2012). Models, fictions, and realism: Two packages. *Philosophy of Science*, *79*(5) (November 19), 738–748.

McClelland, J. L., Rumelhart, D. E. & PDP Research Group (Eds.) (1986). *Parallel Distributed Processing: Explorations in the microstructures of cognition, Vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Montebelli, A., Lowe, R., Ieropoulos, I., Melhuish, C., Greenman, J., & Ziemke, T. (2010). Microbial fuel cell driven behavioral dynamics in robot simulations. In H. Fellermann et al. (Eds.), *Artificial Life XII: Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems* (pp. 749–756). Cambridge, MA: MIT Press. Available at https://mitp-web2.mit.edu/sites/default/files/titles/alife/0262290758chap133.pdf [Accessed November 10, 2011].

Neisser, U. (1963). The imitation of man by machine: The view that machines will think as man does reveals misunderstanding of the nature of human thought. *Science*, *139*(3551): 193–197.

Newell, A., Shaw, J. C., & Simon, H. A. (1960). A variety of intelligent learning in a general problem solver. In M. C. Yovits & S. Cameron (Eds.), *Self-organizing systems: Proceedings of an interdisciplinary conference* (pp. 153–189). Oxford: Pergamon Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(4) (July), 883–99.

Di Paolo, E., Noble, J., & Bullock, S. (2000). Simulation models as opaque thought experiments. In M. Bedau, J. McCaskill, N. Packard & S. Rasmussen (Eds.), *The Seventh International Conference on Artificial Life* (pp. 497–506). Cambridge, MA: MIT Press.

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3) (March 11), 283–311. doi:10.1007/s11229-011-9898-4.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *23*, 73–193.

Santos, D., Sangbae, K., Spenko, M., Parness, A., & Cutkosky, M. (2007). Directional adhesive structures for controlled climbing on smooth vertical surfaces. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 1262–1267. IEEE.

Sanz, R., & Hernández, C. (2010). Autonomy, intelligence and animat mesmerization. In C. Hernández, J. Gómez & R. Sanz (Eds.), *From brains to systems: Preprints of the BICS 2010 conference on brain-inspired cognitive systems* (pp. 256–270). Madrid. Preprint available at http://tierra.aslab.upm.es/events/BIC2010/documents/BICS-2010-Preprints-complete.pdf.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74* (11), 1–29.

Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science*, *17*(3) (October 1), 225–244. doi:10.1080/0269859032000169442.

Suárez, M., (Ed.) (2009). *Fictions in science: Philosophical essays on modeling and idealization*. Vol. 4. New York: Routledge.

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(2), 124–140.

Tolman, E. C. (1939). Prediction of vicarious trial and error by means of the schematic sowbug. *Psychological Review*, *46*(4), 318–336.

Webb, B. (2009). Animals versus animats: Or why not model the real iguana? *Adaptive Behavior*, *17*(4) (July 28), 269–286.

Weber, B. H., & Depew, D. J. (Eds.) (2003). *Evolution and learning: The Baldwin effect reconsidered*. Cambridge, MA: MIT Press.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. New York: Oxford University Press.

Zeigler, B. (1976). *Theory of modelling and simulation*. New York: Wiley.