Synthese (2009) 169:521–538
DOI 10.1007/s11229-008-9432-5

Learning from the existence of models: On psychic machines, tortoises, and computer simulations

Dirk Schlimm

Received: 5 December 2007 / Accepted: 17 October 2008 / Published online: 18 November 2008

Abstract Using four examples of models and computer simulations from the history of psychology, I discuss some of the methodological aspects involved in their construction and use, and I illustrate how the *existence* of a model can demonstrate the viability of a hypothesis that had previously been deemed impossible on a priori grounds. This shows a new way in which scientists can learn from models that extends the analysis of Morgan (1999), who has identified the *construction* and *manipulation* of models as those phases in which learning from models takes place.

Keywords Cognitive psychology · Computer simulations · Learning · Models · History of psychology · Scientific practice · Clark L. Hull · William Grey Walter Methodology

1 Introduction

Models of various sorts play an essential role in psychology. Leaving aside the study of animals as models for human beings, four different artificial models that were introduced in the 20th century are presented in this paper: Hull's psychic machines, Grey Walter's tortoises, Newell and Simon's classical symbolic systems, and Rosenblatt's perceptrons with their extension to multi-layered connectionist networks (Sect. 2). The first two are mechanical models, the others are computational models or computer simulations.[1] Mary Morgan has argued convincingly that scientist can learn from models in two distinct phases, namely during the construction of the

[1] The term "computer simulation" is used differently in psychology than in other disciplines (see Sect. 3).

D. Schlimm (✉)
McGill University, 855 Sherbrooke St. W., Montreal, QC, Canada H3A 2T7
e-mail: dirk.schlimm@mcgill.ca

models and afterwards by using them (Morgan 1999).[2] She has illustrated, on the one hand, that setting up an adequate model requires the identification of key components, and that it involves interpretation, conceptualization, simplification, approximation, and integration in various degrees. Using a model, on the other hand, involves representation, autonomous functioning, and manipulation. A closer look at the four models from the history of psychology supports this analysis (Sect. 3.1). Moreover, it reveals that in the construction of a model a researcher can focus on replicating most accurately some given data or behavior, or on getting the most out of a certain set of basic mechanisms that underlie the functioning of the model. These two methodological approaches are commonly referred to as *analytic* and *synthetic*, and they are discussed in Sect. 3.2, with particular attention to the close relationships between analytic computer simulations and theories, and between synthetic models and agent-based models. While the case studies presented in this paper confirm Morgan's claims that researchers can gain valuable insights during the construction of the models as well as through manipulations, each of these models was also employed in arguments that refute claims about necessary conditions for certain types of behavior or in support of claims about the internal mechanisms that produce certain behavior. These two kinds of arguments are related to the methodological distinction between analytic and synthetic models. In all of these arguments for the validity of a particular claim or the viability of an approach, which were often made in direct response to a previously formulated claim of the contrary, the bare fact that any of these models was exhibited was a significant contribution to scientific progress (Sect. 3.3). This shows how scientists can learn from the *existence* of models, and thus extends Morgan's analysis of learning from models. That the above examples are all drawn from the history of psychology should not detract from the fact that this use of the existence of models in scientific argumentation is nevertheless very general.[3]

2 Physical and computational models in psychology

2.1 Hull's psychic machines

In reaction to the strong mechanistic tendencies in late 19th century physiology (e.g., Helmholtz), the early 20th century saw a revival of *vitalism*, in particular in biology, but also in psychology. Driesch, for example, argued forcefully against 'association' and 'mechanics' and in favor of 'soul' and 'entelechy' as fundamental concepts of psychology (Driesch 1925, p. 267). It was against this background that the young American psychologist Clark L. Hull began building and studying machines that were aimed at simulating certain aspects of human behavior. Early in life he had already

[2] This analysis is also mentioned approvingly in Hartmann and Frigg (2005, p. 745).

[3] See, e.g., semantic consistency proofs and the notion of 'proof of concept' (Sect. 3.3).

constructed various mechanical devices and had developed a mechanistic view of the world (Hull 1952). In his "idea books"[4] Hull writes that

> it has struck me more than once that so far as the thinking processes go, a machine could be built which would do every essential thing that the body does (except growth) so far as concerns thinking, etc. And […] to think through the essentials of such a mechanism would probably be the best way of analyzing out the essential requirements of thinking […] (Hull 1962, p. 820; entry dated March 1, 1926; see also p. 839)

A year after he wrote this passage Hull learned about Pavlov's discovery of the phenomenon of the conditioned reflex and he became convinced that this was the fundamental mechanism underlying learning. Together with collaborators Hull designed and implemented in the course of the next four years mechanical devices to simulate the behavior observed by Pavlov. They experimented with various realizations until they settled for an electric circuit with "polarizable cells and mercury–toluene regulators" (Hull and Baernstein 1929).[5]

At the outset of a classical Pavlovian experiment on conditioning, presenting food to a dog causes it to salivate (*reflex*), whereas the sound of a bell has no such effects. However, after having been presented repeatedly with both food and the sound of a bell at the same time (*conditioning*), the sound of the bell alone suffices to cause the dog to salivate (*conditioned reflex*). Hull simulated such series of events with a machine that had two switches as inputs and a light bulb as output. In the initial state, only one of the switches caused the light to turn on (*reflex*), while switching on the other had no visible effect whatsoever. But, turning on both switches charged the polarizable cell, which could then later be discharged to the light bulb by turning on only the second switch. Thus, after having repeatedly switched on both switches simultaneously (*conditioning*), also the second switch, if turned on by itself, caused the light bulb to glow (*conditioned reflex*). Further aspects of Pavlov's experiments were also reproduced by Hull's device, e.g., that the strength of the conditioned reflex depends on the number of simultaneous stimulations of the inputs, and that the conditioned reflexes decay if the stimuli are not presented together for a longer period of time.

In their presentation of the above results Hull and Baernstein emphasize two points. First, that the existence of such a machine shows that "mental processes are independent of the material substance," and second, that to build such a machine one has to identify the "essential functions" of the behavior that is being modeled. The second point fits exactly Morgan's observation that the building of a model can lead to important insights (Morgan 1999), while the first point is an example of how the existence of a model can teach us something about the necessary and sufficient conditions for certain behavior. More on this later.

[4] Hull kept extensive notebooks, which he called "idea books" and in which he recorded his thoughts and research ideas. From the days of his graduate studies in 1916 to the end of his life in 1952 he completed at least one such book every year, 73 in total. Passages of these books have been published, with an introduction, in Hull (1962).

[5] See also Baernstein and Hull (1931) and Krueger and Hull (1931); for a detailed analysis of Hull's machines, see Cordeschi (1991).

For being able to compare human behavior with that of the model, certain parts of the organism, e.g., sense organs, responding system, and nervous system, must be represented by corresponding components of the model, but Hull refuses to make the additional claim that the underlying mechanisms of the model "are duplicates of the corresponding organic processes" (Baernstein and Hull 1931, p. 99). Indeed, Hull is very careful to point out which of the characteristics of conditioned reflexes that Pavlov had determined experimentally are reproduced by his machines and which are not (e.g., delayed reflexes). For the latter, he expresses the hope that further research with more elaborated machines might eventually lead to their successful simulation. Encouraged by his early successes of imitating very simple learning behavior Hull also envisages the possibility of simulating more complex cognitive functions, such that "at a not very remote date the concept of a 'psychic machine' may become by no means a paradox" (Baernstein and Hull 1931, p. 106)[6] as was the view of the proponents of vitalism.

2.2 Grey Walter's tortoises

Two decades after Hull's work on psychic machines the neurophysiologist William Grey Walter became one of the most well-known builders of mechanical models in Britain.[7] After having achieved groundbreaking results in his research on electroencephalography (EEG), he turned his attention in the early 1950s to the internal workings of the brain. At the time it was common to assume that the brain's performance depends essentially on the number of its neurons. Thus, due the vast number of brain cells, an approach based on modeling seemed to be completely out of reach. To overcome this difficulty, and in stark contrast to the received view, Grey Walter hypothesized that it is not so much the number of units, but the richness of their interconnections that is responsible for generating complex behavior, and he set out to test this claim by building a model with a minimal number of components.

Grey Walter succeeded in devising an autonomous robot with only two functional units, each of which simulated the behavior of a single brain cell. He built two of these machines, which moved on three wheels and consisted only of pairs of tubes, relays, condensers, batteries, and motors (one for moving, the other for steering), and used a photoelectric cell and an electrical contact as inputs. To everybody's surprise these machines were able to exhibit complex and "remarkably unpredictable" animal-like behavior (Grey Walter 1950, p. 44), such as finding a light source in a room and moving towards it while getting around obstacles on their path. These machines were built to continuously explore the environment and Grey Walter referred to them as *Machina speculatrix*, or "tortoises" (Grey Walter 1950). Spurred by the success of these models, Grey Walter went a step further and tried to endow his machines with the ability to learn from previous experiences. Like Hull, he considered Pavlov's conditioned reflex to be the basic mechanism of learning, and he implemented this in a similar type of

[6] See also the entry of 2 July 1930 in Hull's idea books, where he refers to "manuscripts or ideas about the actual design of psychic machines" (Hull 1962, p. 839).

[7] For more background on Grey Walter, see Hayward (2001).

machine, *Machina docilis*, which was also equipped with an additional microphone. Its learning unit, CORA (conditioned reflex analogue), was a small electrical circuit consisting of a handful of amplifying and discharging tubes, condensers, and resistors. At first, the sound of a whistle did not provoke any reaction from the machine, but by repeatedly blowing the whistle at the same time that a light source was shown, the robot would associate the sound and the occurrence of light, such that it eventually became attracted to sound even in the absence of light (Grey Walter 1951). Grey Walter considered his models as genuine tools for scientific inquiry, and he emphasized that they generated unforeseen behavior (e.g., the learning of defensive reflexes) that was nevertheless typical of the animal behavior they were intended to simulate (Grey Walter 1953, pp. 179–181).

2.3 Classical symbolic systems

Soon after digital computers became available at research institutions psychologists realized that these could be used as a new tool for simulating behavior.[8] Thus, in the late 1950s the first theories of cognition were developed that could be implemented as computer programs. In this context the work of Alan Newell and Herbert A. Simon was most influential. In particular, they introduced a new level of analysis of cognitive processes, namely that of (symbolic) "information processes" (Newell et al. 1958). They argued that both computers and human beings can be interpreted as information processing systems, and that the behavior of a particular information processing system is "explained" by a computer program that produces the same behavior.

Thus, to investigate human problem solving behavior at the level of information processes one formulates a computer program and then compares the output generated by the program with the behavior of human subjects. Moreover, since the program is intended to simulate the entire dynamic reasoning process and not just its final outcome, Newell and Simon compared the computer output during various stages of the simulation with verbal thinking-aloud protocols obtained from subjects while they were solving given problems.[9] Finally, if the program was able to simulate human behavior over a wide range of situations Newell and Simon proposed to regard the program itself as "a theory of the behavior" (Newell and Simon 1961a, p. 2012):

> Only when a program simulates the entire sequence of behavior—for example, makes the same chess analysis as the human player—do we have any assurance that we have postulated a set of processes that is sufficient to produce the behavior in question. (Newell and Simon 1961a, p. 2016)

[8] In fact, the computer itself became a popular model for the organization of the brain (e.g., von Neumann 1958).

[9] The following is an excerpt of such a protocol, where the task was to transform a logical expression into another using a set of given rules (clarifying questions from the experimenter are in italics): "I'm looking at the idea of reversing these two things now. *Thinking about reversing what?* The *R*'s … then I'd have a similar group at the beginning but that seems to be … I could easily leave something like that 'til the end, except then I'll…*Applying what rule?* Applying, … for instance, 2. That would require a sign change. *Try to keep talking if you can.* Well … then I look down at rule 3 and that doesn't look any too practical" (Newell and Simon 1961a, Fig. 3).

Given the fact that their program, called the *General Problem Solver*, had fared quite well in imitating how subjects solved various logic problems (Newell and Simon 1961b), they took this as a validation of the fundamental assumption underlying their approach, i.e., that it "provides an unequivocal demonstration that a mechanism can solve problems by functional reasoning" (Newell and Simon 1961a, p. 2014). Later, convinced by a large number of successful simulations, they formulated their famous *physical symbol system hypothesis*, namely that "the necessary and sufficient condition for a system to be capable of thinking" is that it is able to perform certain symbolic processes (Simon 1993, p. 640).[10]

2.4 Perceptrons and neural networks

At the same time when Newell and Simon were analyzing cognitive processes at the information processing level, a radically different approach emerged that took recent findings in neuroscience about the internal workings of the brain as its starting point. Here, highly idealized analogues of neurons and their interconnections are modeled as *neural networks* (or, without the emphasis on the physiological analogy, as *connectionist systems*) consisting of interconnected layers of nodes, each of which having a number of input connections and a single output connection. Depending on the values of the inputs, the output can either be activated or not, thus imitating the firing of a neuron. Due to their close structural similarity to parts of the brain, neural networks have been considered to offer "a reasonable basis for modeling cognitive processes in general" (Rumelhart and McClelland 1986b, p. 110).

The first models of this kind were put forward by Frank Rosenblatt, who named them *perceptrons*. He aimed at "investigating the physical structures and neurodynamic principles which underlie 'natural intelligence.' " (Rosenblatt 1962, pp. v–vi). He considered perceptrons to be *brain models*, by which he meant "any theoretical system which attempts to explain the psychological functioning of a brain in terms of known laws of physics and mathematics, and known facts of neuroanatomy and physiology" (Rosenblatt 1962, p. 3). Perceptrons consist of only three layers of nodes: input, hidden, and output units. Using Rosenblatt's 'perceptron convergence procedure' to update the connection strengths between nodes, a perceptron can be trained to associate inputs with certain desired outputs. Moreover, Rosenblatt proved that if the input–output relation could be learned at all by a perceptron, then this algorithm would eventually yield the necessary connection strengths. Although perceptrons were originally developed as theoretical models, Rosenblatt held that "a brain model may actually be constructed, in physical form, as an aid to determining its logical potentialities and performance" (Rosenblatt 1962, p. 3), without this being one of their essential features, however. He investigated both physical models ("hardware systems") and computer simulations ("digital simulations") of perceptrons himself for testing and comparing their behavior, since mathematical analyses of the more complex systems were lacking. Thus, for Rosenblatt physical models and computer simulations are on par from a methodological point of view, differing only in regard to practical

[10] For the classic exposition of this hypothesis, see Newell and Simon (1972).

matters: he notes that in comparison with hardware systems, computer simulations are more versatile, but much slower.

In 1969 Minsky and Papert were able to prove that the tasks that the single-layered perceptrons could learn belonged only to a restricted class, the 'linearly separable problems'. Furthermore, although they did not rule out in principle that Rosenblatt's learning algorithm could be extended to more complex networks, they contended that "[t]here is no reason to suppose that any of these virtues [of perceptrons] carry over to the many-layered version" (Minsky and Papert 1969, p. 232). In the wake of these results research on connectionist models almost came to a halt and attention in cognitive science was redirected to the problem of knowledge representation. In contrast to the classical computer models, where the computational symbols are claimed to be analogous to mental representations, connectionist models do not have any obvious localizable representation of information. On the one hand this makes them subject to criticisms,[11] but on the other hand this fact itself indicates a further similarity with human brains. It was only in the 1980s with the formulation of the 'generalized delta rule' by Rumelhart and his colleagues that the earlier difficulty was overcome and research on connectionist networks was intensified. Indeed, this has been hailed as "one of the most significant contributions to connectionist research" (Medler 1998, p. 53)[12] and connectionist models have remained to this day an active field of research.

3 Learning from the existence of models

Some general features that are prominent in the above case studies from the history of psychology are discussed next, and we shall see that representational capacity and autonomous functioning enable these models to be used as genuine tools for scientific inquiry (Sect. 3.1). For the construction of models psychologists have developed two methodologies which differ in their focus on either the overall behavior or on the internal mechanisms that generate this behavior (Sect. 3.2). However, despite the differences between the various models and computer simulations under consideration, we shall see that they have played very similar roles in scientific arguments (the distinction between analytic and synthetic models affecting the particular forms of these arguments). I will show how the *existence* of a model can be used to refute necessity claims and to demonstrate the viability of research programmes (Sect. 3.3), which extends the analysis of Morgan, who has identified the *construction* and *manipulation* of models as those phases in which learning from models takes place (Morgan 1999).

3.1 Representational capacity and functional autonomy

Hull, Grey Walter, Newell and Simon, and Rosenblatt all emphasize the predictive power of their models, i.e., their ability to generate unforeseen behavior that the researcher can exploit to formulate novel hypotheses. This is possible because both

[11] See, for example, the debate on compositionality: Fodor and Pylyshyn (1988) and replies.

[12] For a different perspective, see "Prologue: A View from 1988" and "Epilogue: The New Connectionism" in Minsky and Papert (1988, pp. viii–xv and pp. 247–280).

the physical and computational models have representational capacities and function autonomously.

Representational capacity is an essential feature of models, since, in order to interpret the model *as* a model of something, it has to latch on to either theory or the world, i.e., certain features of the model must represent aspects of what it is intended to model. In the case of Grey Walter's tortoises, for example, the photoelectric cell corresponds to a sensory organ such as a moth's eye, and the robot's wheels correspond to means of motions such as a moth's wings. These positive analogues[13] allow us to compare the behavior of the mechanical tortoise to that of a real moth and to conclude that they resemble each other in the sense that both are attracted by a light source. Hull's machines are more primitive in this regard, using switches as input and a light bulb as output. In the case of computational models the representational units are usually the various means by which the computer program receives external input and communicates its output. However, also internal states can serve a representational function. For example, the state in which Newell and Simon's program tries to apply a certain symbolic rule is interpreted as representing the quest of a particular subject to apply a corresponding logical inference. Indeed, the question of the adequateness of knowledge representation in terms of symbolic systems and neural networks has been the source of a long and still unsettled debate between the proponents of the different computational models.

Using the pendulum model and Prandtl's model of a fluid with a bounding layer as case studies, Morrison argued convincingly for a hybrid nature of these models, "neither theory nor simple descriptions of the world" (Morrison 1999, p. 45), which gives rise to their functional independence. This independence in turn forms the basis for their role as *autonomous agents* of scientific inquiry. In other words, Morrison has shown that, despite having significant connections to theory, models are independent sources of scientific knowledge. The functioning of Hull's and Grey Walter's mechanical models goes beyond the influence of the scientist and his theory, despite their being constructed with reliance to theoretical considerations. This autonomy of the models is the basis for their being able to behave in unforeseen ways, such as displaying particular aspects of conditioned learning that were not purposely built into them by Hull and his collaborators, and exhibiting interesting interactive dynamics between more than one of Gray Walter's tortoises. Similarly, although the behavior of the computer models is determined only by the program and the internal logic of the computer and is thus deterministic, it still is outside the complete control of the modelers since they are not omniscient with regard to deductive consequences. Indeed, researchers that devise computer simulations often refer to their work as *experiments* and thereby emphasize the autonomy of their models. Thus, all of the physical models and computer simulations discussed above exemplify Morrison's observation about the functional autonomy of models.[14] This, and their representational capacity enable us to use them as genuine tools for scientific discovery.

[13] See Hesse (1966).

[14] See also Morrison and Morgan (1999), where this autonomy is employed to characterize models as *instruments*.

3.2 Analytic and synthetic models

The construction and use of models can focus on either the internal mechanisms or on the overall behavior of the model. This distinction also plays a role in how the existence of models is used in arguments about the viability of particular research programmes, which addressed in the next section. The methodological distinction between analytic and synthetic models[15] is orthogonal to that between physical and computational models. For example, there are striking similarities between Hull's use of mechanical models of animal behavior of the 1930s and Newell and Simon's research based on computer programs that simulate cognitive processes of the 1960s, despite the obvious technical differences between them.[16] Both approaches take certain data that is to be reproduced (animal behavior and verbal protocols of problem solving) as their starting points, which is the characteristic feature of *analytic*, or data-driven, models. In this connection it is interesting to notice that the methodological considerations formulated by Hull and by Newell and Simon are indeed very similar, in particular their expressed agnosticism about whether or not the particular implementations of their models exactly replicate the mechanisms that generate the behavior that is simulated. An alternative methodological approach, which is exemplified by Grey Walter's robots and by research on neural networks, is referred to as *synthetic*. Here, the researchers take certain basic building blocks, whose functioning is well understood, and their configurations as the starting point for the construction of models.[17] Thus, we get the following classification of the discussed models:

	Physical	Computational
Analytic	Hull's psychic machines	Newell and Simon's symbolic systems
Synthetic	Grey Walter's tortoises	Perceptrons and neural networks

3.2.1 Complexity and understanding

The behavior of an agent typically depends in part on its internal mechanisms and in part on the environment. A researcher who wants to analyze the agent's behavior is faced with the problem of determining exactly how much of it is due to the internal structure. Unfortunately, experience has shown that "when we analyze a mechanism, we ted to overestimate its complexity" (Braitenberg 1984, p. 20).[18] Grey Walter's tortoises provide a compelling illustration of this claim, since, when they were exhibited for the very first time in public, the audience was extremely surprised after it

[15] See Dawson (2004), in particular pp. 3 and 98–100 for a discussion of motivations for these approaches.

[16] Incidentally, one of the other psychologists who showed an interest in the use of computers for modelling cognitive processes in the 1960s was Carl I. Hovland (Hovland 1960), who had collaborated two decades earlier with Hull (Hull et al. 1940).

[17] One also finds the term "synthetic models" to be used for artificial models in general, but this is not the sense intended here.

[18] Dawson refers to this observation as the "law of uphill analysis and downhill synthesis" (Dawson 2004, p. 3).

was informed about the internal simplicity of the robots. This observation explains a general difficulty that analytic models (and theories) of behavior face, namely, that in order to account for a wide range of behavior they tend to become quickly very complex. Indeed, it is often the case that each new aspect of behavior that is to be simulated leads to an ad hoc extension of the current model. An illustration of this is provided by historical development of Hull's theory of adaptive behavior, which he turned to after his work on psychic machines, and which "just broke down in its enormous detail, with all the exceptions to it" (Baars 1986, p. 113). (This commonality between analytic models and theories is taken up again below.)

This methodological difficulty of analytic models is overcome, at least prima facie, by synthetic models, since they are built from very simple and well-understood parts (e.g., simple electronic components and connectionist units) and thus promise to lead to simpler theories. Nevertheless, complexity is often introduced through the back door, because often very many of these building blocks must be organized into a single system (e.g., a neural network), whose overall behavior is again difficult to analyze despite the fact that the behavior of the components is easily understood. In the end, computational models, be they analytic or synthetic, can become as complex as the process they are trying to simulate, thus yielding no immediate advance in understanding.[19] However, while this difficulty has serious implications on the epistemic value of these models, it only bears very little on how we can learn from the existence of models, which is discussed in the next section.

3.2.2 Analytic simulations as theories

It is a noteworthy observation that the use of analytic computer simulations in cognitive psychology blurs the distinction between models and theories. Computer simulations are often described in the literature as being based on an underlying well-developed theory.[20] However, there was no such underlying theory that supported the introduction of classical symbolic models in psychology. Instead, the simulations themselves have been regarded as theories, for example, by Newell and Simon, who explicitly draw an analogy between theories that are expressed by computer programs and by mathematical equations, and they attribute the same epistemological status to computer programs and mathematical formulas (Newell and Simon 1961a, p. 2013). In fact, their methodology is in accord with that of mathematical modeling in psychology, where the models result from cycles of theory construction from given data, deduction of consequences by logical and mathematical operations, and verification or refutation on the basis of newly collected data (Dawson 2004, p. 35).

[19] This well-known problem is also referred to as *Bonini's Paradox* (Dutton and Briggs 1971, p. 103); it is discussed, for example, in Churchland and Sejnowski (1988), Lewandowsky (1993), and Dawson (2004, pp. 17–18).

[20] See, e.g., Humphreys (2004, pp. 107–108) and Hartmann and Frigg (2005, p. 745). Also Morrison mentions the common view that model building starts with a background theory in Morrison (1999, p. 51).

3.2.3 Synthetic models in other sciences

The synthetic models of the psychologists are closely related to the *agent-based* or *individual-based models* that are mainly used in the social sciences, where the behavior of an entire population is modeled via interactions of its individuals. Schelling's *tipping model* is a well-known early example, where an artificial society is modeled in which an agent would stay put if more than one third of its neighbors are of the same color and would move to another location on a two-dimensional grid otherwise. These simulations show that even though each agent is somewhat tolerant segregated neighborhoods are formed the long run (Schelling 1978).[21] In these kinds of models

> there is no overarching model of the social system, only rules governing the interaction of the individuals comprising the system. […] Phenomena often emerge within these models at the macro level that are the result of multiple interaction at the micro level, these macro phenomena being unforeseen when the individual-level analysis begun. (Humphreys 2004, p. 131)

The analogy to neural networks is striking: Each node in the network corresponds to an individual and the entire network corresponds to the population. The lack of an overarching model for the system corresponds to the lack of a well-developed underlying theory, and the ability of networks to recognize input-output patterns emerges from the interactions of the single nodes, whose behavior is clearly specified.

Because of their similarity, the dangers that have been identified for the agent-based modeling approach, like the return to a priori science and the need for a justification of the basic mechanisms that goes beyond their ability to simulate the desired phenomena, carry over to synthetic models. Due to the autonomy of computer simulations and the ease in which they can be developed and tested, it is easily possible to sever the contact with the empirical basis that motivated their development. In particular, synthetic models can be built and studied without any reference to empirical data and completely independently of what they are later be said to be models of. For example, connectionist networks are also studied by computer scientists as an instance of "machine learning" techniques (Mitchell 1997). The synthetic psychologists discussed above were indeed painfully aware of the need of independent justifications of their basic assumptions and they often cited experimental results to show that there are good reasons to hold that the mechanisms they postulated did in fact correspond to genuine features of the systems under investigation. For example, the importance of electricity for brain functions was demonstrated by Grey Walter's own research on EEGs, and the functioning of connectionist units was explicitly motivated by independent findings in neurophysiology.[22]

[21] For an overview and discussion of these kinds of models, see Axelrod (1997).

[22] See Grey Walter (1950, pp. 42–43), Rosenblatt (1962, p. 3), and Rumelhart and McClelland (1986b, p. 110).

3.3 Existence of models in scientific argumentation

In the early development of a scientific discipline one regularly encounters claims about certain necessary conditions and about what is possible or not, of the form "H is a necessary condition for G," "it is impossible to obtain G without H," or "$\neg H$ cannot cause G." In the case of psychology, claims of this kind have been: "*vis viva* is a necessary condition for intelligent behavior,"[23] "it is impossible to have intelligent behavior by purely mechanical means," "symbolic processes alone cannot generate intelligent behavior," "complex mechanisms are necessary to generate complex behavior," and "connectionist networks can only accomplish very simple tasks." For H to be necessary for G is logically equivalent to the claim that whenever you have G, you also have H, or, formally, $\forall x\,(Gx \rightarrow Hx)$.[24] To refute such an assertion one has to exhibit an instance of Gx that is not also an instance of Hx,[25] in other words, an instance, say a, that makes $Ga \,\&\, \neg Ha$ true. Incidentally, such an instance also establishes that Gx and $\neg Hx$ are consistent with each other. Under the assumption that 'purely mechanical' is the contrary of 'containing some form of vital force', a device a that exhibits some intelligent behavior (i.e., Ga holds) and is purely mechanical (i.e., $\neg Ha$ holds), thus, refutes the claim that 'to contain some form of vital force is necessary to exhibit some form of intelligent behavior' (i.e., that Hx is necessary for Gx). In other words, exhibiting the *existence* of a model of a certain kind is all that is required to falsify a necessity claim of the form under consideration, and it is precisely in this sense that scientists can learn something just from the bare existence of models.

The move from talk about models, be they physical, computational, or otherwise, to talk about models in a *logical* sense, i.e., that interpret the primitive terms of a language and that can satisfy statements of that language, should not be taken as an endorsement of Suppes' claim that the logical notion of model "is the fundamental one for the empirical sciences as well as in mathematics" (Suppes 1962, p. 252). Rather, for the sake of this discussion we only need to accept that any model *can* be understood as a model in the logical sense, where the notion of a logical model is suitably extended to include physical objects. Since every model has certain representational capacities, we can easily consider these to interpret particular terms in our language. For example, the term 'auditory organ' is so interpreted that the microphone of Grey Walter's tortoise counts as one.

These considerations are closely related to the use of models for semantic consistency proofs and independence proofs in mathematical practice. Here models, usually conceived as abstract mathematical structures, are also understood to interpret the primitive terms of a theory. In particular, if a model for a theory can be exhibited one can conclude that the theory is consistent, i.e., that it cannot lead to contradictions, since otherwise these contradictions would be reflected in the model, too. Thus, the bare fact that a model exists can constitute an important advance also in mathematics. Famous examples of this use of models are the models for non-Euclidean geometries

[23] 'Intelligent' is meant here to include typical animal-like and human-like behavior.

[24] For the analysis of these simple forms of argument, the use of modal logic can be dispensed with.

[25] To emphasize that G and H are predicated of something, I shall write Gx and Hx from now on.

put forward by Beltrami and Klein, which established beyond doubt that the geometries investigated by Bolyai and Lobachevsky were consistent (in other words, that Euclid's parallel postulate is independent of his other axioms) and thus of genuine mathematical interest.[26]

Let us now revisit the four models presented above with an eye to the way in which their existence was used in arguments about what is possible and what is not. Hull's "idea books" show him becoming more and more convinced of the mechanical nature of cognitive processes, e.g., he talks about the "human machine" in 1925, refers to his own views as "mechanistic psychology" in 1927, and mentions "psychic machines" in 1930.[27] At the same time, he realized that his own development had been held back "probably ten or fifteen years at least" by the wide-spread "dogma that an organism made up of consciousless particles may not possibly manifest consciousness" and he lamented that "the world is so hypnotized by the ancient animism" (June 16, 1930; Hull 1962, pp. 837–838). To overcome these views, Hull speculated that he might be most successful "especially if I construct a series of striking psychic machines to support the theory" (February 26, 1930; Hull 1962, p. 833). Finally, Hull explicitly replied in print to those who argued that behavior, which is characterized as involving a psyche or being intelligent, cannot be generated by purely mechanical means in his closing remarks of Krueger and Hull (1931), in which he presents an electromechanical model that imitates Pavlov's conditioned reflex without any recourse to any 'psychic' forces. He positions his work in direct opposition to the "very widespread and persistent [belief] that certain complex forms of adaptation cannot take place by any imaginable concatenation of materials without the mediation of some *nous*, *entelechy*, soul, spirit, ego, mind, consciousness, or *Einsicht*" (Krueger and Hull 1931, p. 267). Thus, since his mechanical model does simulate certain aspects of the behavior in question, it refutes the claim that any of these notions is necessary for producing the behavior that is simulated. Moreover, since Hull's aim was to reproduce certain patterns of learning behavior, rather than arguing in favor of some particular underlying mechanism that generates this behavior, his methodology is analytic and his refutation does not depend on the specific implementation of the model he put forward. It is interesting to note that Hull also anticipated that the more successful the proponents of the mechanistic psychology will be, i.e., the more types of behavior they will succeed in generating by mechanical means, the proponents of vitalism "will gradually retreat to more and more inaccessible parts of the psychological terrain" (Krueger and Hull 1931, p. 267), and the later development of cognitive psychology has vindicated this prediction.

With the help of computers Newell and Simon were able to show that even more complex forms of behavior, such as solving logical problems, could also be simulated by mechanical means alone. Like Hull, they also emphasize the mechanical character of their model and direct this against proponents of vitalism, e.g., in the opening remarks of "Computer simulation of human thinking and problem solving": "It is no longer necessary to argue that computers can be used to simulate human

[26] For more on the history of these models, see Bonola (1955) and Gray (2004).

[27] See Hull (1962, pp. 820, 823, 828); see also the quotation from 1926 in Sect. 2.1 above.

thinking," since, by the various computer simulations "the proof of possibility" of a purely mechanical simulation of cognitive processes has been accomplished (Simon and Newell 1961, p. 137). The notion of proving the practical realizability of an idea by providing an actual model is also referred to as "proof of concept" and it has become a well-known technique in various fields of engineering.[28]

As refutations of the claim that some form of vital force is necessary to produce certain patterns of behavior, both Hull's and Newell and Simon's models only had to be purely mechanical, regardless of the particular mechanisms they were based upon. This is different in case of the arguments of Grey Walter and the proponents of connectionist networks. The claim that Grey Walter set out to refute with his autonomous robots was the widely held belief that complex animal-like behavior must be based on a large number of internal components. A early mechanical model that supported this view was W. R. Ashby's *homeostat*, which had a large number of different internal states but exhibited only very rudimentary behavior, "like a fireside cat or dog which only stirs when disturbed, and then methodically finds a comfortable position and goes to sleep again" (Grey Walter 1953, p. 123).[29] In direct opposition to this received view, Grey Walter hypothesized that "the elaboration of cerebral functions may possibly derive not so much from the number of its units, as from *the richness of their interconnection*," and he noticed that "this speculation had the great advantage that its validity could be tested experimentally" (Grey Walter 1953, p. 118; emphasis in original). Such a test would consist in the construction of a "working model that would behave like a very simple animal" with a minimal number of components (Grey Walter 1953, p. 125). Thus, by building his tortoises that consisted of only a very small number of internal components, but that exhibited rather complicated patterns of behavior, Grey Walter was able to demonstrate the correctness of his hypothesis. The fact that the focus of Grey Walter's argument is on the mechanisms that generate the behavior and not primarily on the behavior itself, explains his commitment for a synthetic methodology for the construction of his models. This also holds for the computational models of Rosenblatt and Rumelhart.

Rosenblatt's perceptrons, which could be trained a range of different tasks, were the first models that refuted the common view that computers cannot learn associations between inputs and outputs other than those that have been explicitly included in their program. Later, after showing the theoretical limitations of these networks that had only a single hidden layer of nodes, Minsky and Papert speculated that Rosenblatt's learning algorithm could not be extended to multi-layered networks (Minsky and Papert 1969, p. 232). When Rumelhart and his colleagues actually came up with an algorithm that could solve this task they proudly announced: "we believe that we have answered Minsky and Papert's challenge and *have* found a learning result sufficiently powerful to demonstrate that their pessimism about learning in multilayer machines was misplaced" (Rumelhart et al. 1986a, p. 361; emphasis in original). Again, an impossibility claim was refuted by the existence of a model.

[28] See, e.g., Weston (2004, Chap. 7).

[29] Indeed, Ashby's model "could be interpreted as supporting the claim that the complexity of the behavior of whole organisms largely emerges from (1) a large number of internal components, and (2) the interactions between these components" (Dawson 2004, p. 83).

One might object that the above questions were not settled definitively by the models that were put forward, as some of them are still topics of current debates. This leads to an important issue in the discussion of the roles of models in science, namely the question of whether something is indeed a model for what it is claimed to model. In the present discussion the criteria of adequacy boil down to the question whether the goal property *G* is in fact characteristic for the notion that it is intended to capture. Hull's model has shown that an electro-mechanical device can be built that exhibits behavior which is analogous to Pavlov's conditioned reflex. However, the question whether that is all there is to learning is thereby not addressed. Even more problematic has been whether solving logic problems counts as genuinely intelligent behavior, as Newell and Simon contended. Indeed, opponents to the view that computers can exhibit intelligent behavior can always retreat to the (admittedly dubious) position that if behavior has been simulated by a computer it cannot be intelligent behavior.[30] A debate of a different kind is whether physical symbol systems or connectionist models are the best way to study intelligence. But, also in this debate the presence or absence of particular models has also often been used to support one position or the other. To illustrate: On the one hand, Simon uses the lack of concrete simulations of complex cognitive performances as an argument against neural networks as models for human thinking (Simon 1993, p. 640), while researchers on neural networks, on the other hand, are keen to meet challenges of this kind, by demonstrating that their models do indeed simulate aspects of human behavior (e.g., Shultz 2003, pp. 221–250).

The fact that the models and simulations did succeed to mirror some aspects of human behavior is also taken by the modelers as providing some information with regard to the mechanisms that produce the behavior in question. In general, by providing models (mechanical or computational) of certain phenomena, a step is made towards uncovering the underlying mechanisms. Thus, in a sense the phenomena have been stripped from the veil of mysteriousness that had covered them. This demystifying role of models is a theme that is often repeated by cognitive psychologists. Hull, for example, writes that

> [i]t is believed that the construction and study of models of the type described above will aid in freeing the science of complex adaptive mammalian behavior from the mysticism which ever haunts it. (Krueger and Hull 1931, p. 267)

In a similar vein, Newell and Simon consider their computational model as

> a good approximation to an information-processing theory of certain kinds of thinking and problem-solving behavior. The process of thinking can no longer be regarded as completely mysterious. (Newell and Simon 1961a, p. 2016)

The demystification of consciousness through the building of machines is mentioned again three decades later by the neuroscientist Francis Crick:

> If we could build machines that had these astonishing characteristics [of the brain], and could follow exactly how they worked, we might find it easier to

[30] See Hull's anticipation of this move in Krueger and Hull (1931, p. 267). An excellent overview of the debate whether computers can exhibit intelligence or not can be found in Dreyfus (1992).

> grasp the workings of the human brain. The mysterious aspects of consciousness might disappear, just as the mysterious aspects of embryology have largely disappeared now that we know about the capabilities of DNA, RNA, and protein. (Crick 1994, pp. 256–257)

Finally, a word of caution is appropriate regarding the limits of what can be established by the existence of models. Having refuted an alleged claim of necessity by exhibiting a model a that satisfies $Gx \,\&\, \neg Hx$ also entitles one to the claim of having established the statement $Ha \rightarrow Ga$. This means that for the instance a, the property of being a H is *sufficient* for also being a G, but not that this is the case in general, i.e., that $\forall x\,(Hx \rightarrow Gx)$. In particular, it could be that the property H has nothing to do with G (this is related to the *problem of irrelevance* in the context of Hempel's D-N account of explanation). Moreover, having shown that the necessity claim $\forall x\,(Gx \rightarrow Hx)$ is false by exhibiting a model for $Gx \,\&\, \neg Hx$, does not in the least amount to the converse necessity claim $\forall x\,(Gx \rightarrow \neg Hx)$, i.e., that $\neg H$ is *necessary* for G. This is why Newell and Simon formulate their later view that the ability of performing certain symbolic processes is sufficient and necessary for thinking, which is motivated, but not conclusively established by their work, only as an hypothesis (Simon 1993, p. 640).

4 Conclusion

I this paper I illustrated by means of examples of classical and connectionist computer simulations and of two earlier mechanical models that psychologists have learned certain lessons (e.g., that learning is independent of vital forces, that few simple components can generate complex behavior, that symbolic processing can imitate the problem solving behavior of human beings, and that networks formed of very simple building blocks can be trained to solve complex tasks) from the bare existence of these models. In this respect physical models, computer simulations, and even purely theoretical models can perform an important function in the quest for scientific knowledge, namely to demonstrate the viability of a particular approach, to validate or refute certain hypotheses, and to demystify a domain of inquiry. In addition, I have shown a close connection between these arguments and the particular methodology (analytic or synthetic) that is adhered to in the construction of the respective models. Thus, by taking a closer look at the history and the practice of modeling in psychology, novel aspects of the use of models and computer simulations in scientific practice have been brought to light.

Acknowledgements Presented at *Models and Simulations*, Paris, June 12–13, 2006. I would like to thank Uljana Feest, Michael Hallett, Brian van den Broek, and an anonymous referee for many valuable comments on this paper.

References

Axelrod, R. (1997). *The complexity of cooperation. Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Baars, B. J. (1986). *The cognitive revolution in psychology*. New York: The Guilford Press.

Baernstein, H. D., & Hull, C. L. (1931). A mechanical parallel to the conditioned reflex. *Journal of General Psychology, 5*, 99–106.
Bonola, R. (1955). *Non-euclidean geometry: A critical and historical study of its development*. New York: Dover.
Braitenberg, V. (1984). *Vehicles: Explorations in synthetic psychology*. Cambridge, MA: MIT.
Churchland, P. S., & Sejnowski, T. (1988). Perspectives on cognitive neuroscience. *Science, 242*(4879), 741–745.
Cordeschi, R. (1991). The discovery of the artificial. Some protocybernetic developments 1930–1940. *AI and Society, 5*, 218–238.
Crick, F. (1994). *The astonishing hypothesis. The scientific search for the soul*. New York, NY: Touchstone.
Dawson, M. R. W. (2004). *Minds and machines. Connectionism and psychological modeling*. Oxford: Blackwell.
Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT.
Driesch, H. (1925). *The crisis in psychology*. Princeton, NJ: Princeton University Press.
Dutton, J. M., & Briggs, W. G. (1971). Simulation model construction. In J. M. Dutton & W. H. Starbuck (Eds.), *Computer simulation of human behavior* (pp. 103–126). New York: Wiley.
Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.
Gray, J. J. (2004). *János Bolyai, non-Euclidean geometry and the nature of space*. Cambridge, MA: MIT.
Grey Walter, W. (1950). An imitation of life. *Scientific American, 182*(5), 42–45.
Grey Walter, W. (1951). A machine that learns. *Scientific American, 184*(8), 60–63.
Grey Walter, W. (1953). *The living brain*. New York, NY: Norton & Co.
Hartmann, S., & Frigg, R. (2005). Scientific models. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An Encyclopedia* (Vol. 2, pp. 740–749). London: Routledge.
Hayward, R. (2001). The tortoise and the love-machine: Grey Walter and the politics of electroencephalography. *Science in Context, 14*(4), 615–641.
Hesse, M. B. (1966). *Models and analogies in science*. Notre Dame, IN: University of Notre Dame Press.
Hovland, C. I. (1960). Computer simulation of thinking. *American Psychologist, 15*, 687–693.
Hull, C. L. (1952). Autobiography. In E. G. Boring, H. S. Langfeld, H. Werner, & R. M. Yerkes (Eds.), *A history of psychology in autobiography* (Vol. IV, pp. 143–162). Worcester, MA: Clark University Press.
Hull, C. L. (1962). Psychology of the scientist: IV. Passages from the 'idea books' of Clark L. Hull. *Perceptual and Motor Skills, 15*, 806–882.
Hull, C. L., & Baernstein, H. D. (1929). A mechanical parallel to the conditioned reflex. *Science, 70*(1801), 14–15.
Hull, C. L., Hovland, C. I., Ross, R. T., Hall, M., Perkins, D. T., & Fitch, F. B. (1940). *Mathematico-deductive theory of rote learning. A study in scientific methodology*. New Haven, CT: Yale University Press.
Humphreys, P. (2004). *Extending ourselves. Computational science, empiricism, and scientific method*. Oxford: Oxford University Press.
Krueger, R. G., & Hull, C. L. (1931). An electro-chemical parallel to the conditioned reflex. *Journal of General Psychology, 5*, 262–269.
Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4*(4), 236–243.
Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys, 1*(2):18–72.
Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT.
Minsky, M., & Papert, S. (1988). *Perceptrons* (expanded edition). Cambridge, MA: MIT.
Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Morgan, M. S. (1999). Learning from models. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators* (pp. 347–388). Cambridge: Cambridge University Press.
Morrison, M. (1999). Models as autonomous agents. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators* (pp. 38–65). Cambridge: Cambridge University Press.
Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators* (pp. 10–37). Cambridge: Cambridge University Press.
Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*(3), 151–166.
Newell, A., & Simon, H. A. (1961a). Computer simulation of human thinking. *Science, 134*(3495), 2011–2017.

Newell, A., & Simon, H. A. (1961b). Simulation of human thought. In D. Wayne (Ed.), *Current trends in psychological theory* (pp. 152–179). Pittsburgh, PA: University of Pittsburgh Press.

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

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington DC: Spartan Books.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing* (Chap. 8, pp. 318–362). Cambridge, MA: MIT.

Rumelhart, D. E., & McClelland, J. L. (1986b). PDP Models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing* (pp. 110–146). Cambridge, MA: MIT.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: W.W. Norton.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT.

Simon, H. A. (1993). The human mind: The symbolic level. *Proceedings of the American Philosophical Society, 137*(4), 638–647.

Simon, H. A., & Newell, A. (1961). Computer simulation of human thinking and problem solving. *Monographs of the Society for Research in Child Development, 27*(2), 137–150.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski, (Eds.), *Logic, Methodology and Philosophy of Science. Proceedings of the 1960 International Congress in Stanford, CA* (pp. 252–261). Stanford, CA: Stanford University Press. (Reprinted from *Studies in the methodology and foundations of science. Selected papers from 1951 to 1969*, pp. 24–35, by P. Suppes, 1969, Dodrecht, Holland: D. Reidel.)

von Neumann, J. (1958). *The computer and the brain*. New Haven: Yale University Press.

Weston, P. (2004). *Bioinformatics software engineering*. West Sussex, England: Wiley.