CHAPTER ONE

REVERSE ENGINEERING IN COGNITIVE SCIENCE

MARCIN MIŁKOWSKI

1. Three Flavours of Reverse Engineering

The notion of "reverse engineering" has long been embraced by philosophy of science. For example, Daniel Dennett defended his claim that biology is engineering by pointing to some methods of investigation that bear close resemblance to a specific way of understanding artefacts, namely reverse engineering (Dennett 1995, 212-20). Focus on methods of investigation, in general, is a distinctive feature of naturalistic philosophy of science: it is not interested solely in questions of rational reconstruction and justification of scientific theories. It also reflects upon discovery as part of the way science works. Thus, by saying that sciences use reverse engineering, one commits oneself to investigation of strategies that scientists use in discovering true and important invariant generalizations.

Also, by stressing that technology and science both use engineering, philosophers of science target regularities that help them unify the worlds of disparate disciplines in a single theoretical framework. This kind of theoretical unification may be illuminating for both science and technology (even if it is not fully explanatory as unification in Kitcher's 1989 sense of the term).

But what exactly is reverse engineering?

> Reverse engineering is just what the term implies: the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation (Dennett 1994, 683)

A similar notion of reverse engineering was used by Robert Richardson who defines it as "inferring adaptive function from structure" (Richardson 2003, 1277). Note the addition of "adaptive": the design

considerations are linked with considerations of adaptation. Is this a necessary feature of reverse engineering, or maybe there is a special kind of it, which is prevalent in evolutionary biology, as Richardson seems to suggest?

To answer this question, we might be tempted to look at the usage of the notion in computer science and information technology, the original source of the expression. Alas, the usage is far from consistent. In its most frequent uses, as found in thousands of software licences, it is used to refer to deriving source code (which is explicitly banned). But for some, it is not only deriving the code but doing something with it, for example to circumvent copying restrictions, the latter being also called "reengineering". I will follow the practice of many authors that write about reverse engineering (for example, Eilam & Chikofsky 2005) and use the influential paper from *IEEE Software* that legislated conceptual distinctions between the notions. Reverse engineering was defined there as "the process of analyzing a subject system to identify the system's components and their interrelationships and create representations of the system in another form or at a higher level of abstraction" (Chikofsky & Cross 1990, 15).

It is immediately clear that this is a very broad notion indeed. All it takes for a process to qualify as reverse engineering is to create representations at higher level of abstraction, a task that some understand as the very essence of science, and to analyse the structure of a complex system this way.

Consequently, any mechanistic explanation (Machamer, Darden and Craver 2000) will be supported by reverse engineering in this sense. Even the admittedly broad notion of the mechanism, usually understood as a system that has some system-level functional capacity constituted by the orchestrated activity of component parts of the mechanism, is more stringent than the notion of the system implied by the original definition, as no system-level capacity is ever mentioned. Moreover, there is no mention of function either.

With such a broad notion of reverse engineering, which is conflated, as it seems, with any kind of theorizing about complex systems, it is hardly a surprise that many disciplines of science will turn out to be engaged in reverse engineering. Only if you do not focus on complex systems, say in certain branches of physics, might you be doing something else. But then the claim is not really interesting. If cognitive science is reverse engineering, probably just like any special science, so what?

So maybe the notion of reengineering will be more telling:

> Reengineering, also known as both renovation and reclamation, is the examination and alteration of a subject system to reconstitute it in a new

> form and the subsequent implementation of the new form (Chikofsky & Cross 1990, 15).

This notion is definitely narrower: you need to reconstitute a system in a new form, or replicate it somehow. While many sciences replicate phenomena in various models, this is at least not universal, so the claim that cognitive science uses reengineering is far more substantial.

Where does this leave us? We have now three renderings of the claim that cognitive science is reverse engineering (actually, even more if you care about quantification). First, that cognitive science infers function from structure (or even adaptive function). This is a substantial claim, but one can easily point to numerous examples from evolutionary psychology (see Richardson 2007, chapter 2, on reverse engineering in this sense).

Second, that cognitive science uses decomposition strategies to understand cognitive systems, as many other sciences do (Bechtel & Richardson 1993). This is on the verge of being trivial, and not really worth mentioning, as functionalist decomposition is a methodological strategy prescribed by the mainstream philosophy of cognitive science since its very beginnings. Yet some think that these strategies are invalid as cognitive systems are too complex: their evolved biological complexity is to escape the reductionist strategies of reverse engineering (Schierwagen 2012). Alas, arguments that support this bold claim are pretty weak. Schierwagen draws inductive inference from methodologically unsound attempts to computationally simulate the cortical column to the strong conclusion that all reverse engineering will fail. Also, he appeals to Rosen's (1991) claim that biological complexity cannot be analyzed reductively. He supposes however that reverse engineering requires that the capacities of the whole mechanism be identified with capacities of the parts, and that the mechanisms be aggregative in Wimsatt's (2002) sense. This premise is definitely false, and Bechtel and Richardson explicitly deny it by stressing that aggregative systems are an extreme case (1993, 25).

Third, that cognitive science uses reverse engineering and reengineering to replicate the structure of cognitive systems and understand their function in this way. This is what I will focus on the rest of this paper, as there are specific virtues of reengineering in cognitive science.

I already mentioned that the claim regarding the role of reverse engineering might be quantified in different ways: is all cognitive science reverse engineering or just some? It transpires that on the trivial reading of reverse engineering, as functional decomposition, all or almost all cognitive science would refer to cognitive systems as complex (you do not need to believe in modularity to say that cognitive systems have at least

two component parts). But on more substantial readings, not all research methodologies used in cognitive science will resemble reverse engineering. In particular, traditional information-processing psychology (such as Miller 1956) was not interested in reengineering: replicating or simulating cognition. Simulation is a specific tool in cognitive science, and I do not claim that it is required for or used in all cognitive research.

A simulation in cognitive science is a model which serves as an idealization of the phenomenon under consideration. Simulations have finite precision and cannot be used to predict all the attributes of the modelled phenomenon, yet they must be predictive about some. This means that they are products of reverse engineering in the Chikofsky and Cross sense: they are representations in another form or at a higher level of abstraction, even if they are reimplemented physically in another medium. They are not straightforward copies of the phenomena that they describe. Otherwise, using simulations instead of original phenomena would make no sense: there must be some advantage in building a simulation in the first place. One of these advantages is that simulations involve reduction of information–some of it is discarded as noise. The information however must still be there, and this is why simulation remains representational while idealizing.

In what follows, I discuss whether there are some lessons for philosophical inquiry over the nature of simulation to be learnt from the practical methodology of reengineering. I will argue that reengineering serves a similar purpose as simulations in theoretical science, and that the procedures and heuristics of reengineering help to develop solutions to outstanding problems of simulation.

2. Organizational Invariance

For reengineering to work, it must be possible to replicate the system in question, or the phenomenon to be reconstituted. If replication uses a different medium, the phenomenon must be *organizationally invariant* (Chalmers 2011) so that the copies can be *substrate-neutral* (Dennett 1995, 50). Otherwise, the causal structure of the physical system could not be replicated in another medium, using some other substrate. But organizational invariance or substrate neutrality is not to be confounded with multiple realization. The latter notion is used in different ways, and there are plausible reasons to remain sceptical of many claims traditionally connected with multiple realization, especially when it is used to argue for antireductionism (for such criticism, see Polger 2004 or Shapiro 2000, 2004, 2008).

To see the difference between substrate-neutrality and multiple realization, we need to note that multiple realization requires that a single capacity be realized in multiple ways. But not all physical differences make any difference for realization: the colour of paint on the wind tunnel cannot be used to differentiate realizations. Likewise, who made the mouse trap is irrelevant for its capacity to catch mice. What is crucial is that the functional organization that contributes to the capacity being realized is different. Functional organization is basically the causal structure of the system that has some capacity. When reengineering a capacity, we want to replicate it in a *new* or *different* form. But we can speak of replication only when the causal structure, or a causal model of a capacity is the same or very similar.

To explain the differences between organizational invariance, which is basically retaining the same causal structure or topology in different substrates, and multiple realization, it is useful to introduce a simple example of two different physical implementations of similar computers. We will also see that in an essential way the talk of multiple realization is interest-relative. Let us then look at two very similar computers: IBM 709 and IBM 7090. The latter one was a transistorized version of the first one (this example is taken from Wimsatt 2002). Logically, these computers were equivalent, so one could run the same software on both. In other words, these computers are input-output equivalent on every level of detail of their software: any routine in any program you take will be performed in an equivalent way by IBM 709 and 7090. But they are not completely equivalent, as they perform their functions in a different way, so the causal pathway between the input data and output data is not the same at the electronic level. For example, one machine is slower than another, and transistors break in different ways than tubes. The question is whether IBM 709 and 7090 are different realizations of the same capacity. While it is quite clear that the capacity to execute the software of IBM 709 is substrate neutral (it might be emulated on any modern machine as well), it is not so obvious that its realization is different in the two machines in question. For one, the *relevant* causal organization must be the same for them to run the same software. If we conceive the capacity as executing-the-machine-code-and-interfacing-the-peripherals, then causal models of it in both machines will be the same: the differences of speed and breakdown patterns are as inessential as the paint on the wind tunnel. They make no difference to *this* capacity. The organization stays invariant. Yet if we include the speed and breakdown patterns, say, in the specification of the capacity, then the model of the causal structure will include information about electronic elements as well. Otherwise, we could not account for

differences in speed. Both computers, however, have then *different* capacities, so it is no longer true that it is the same capacity realized in different ways.

A traditional proponent of multiple realization might reply that I need not distinguish between substrate-neutrality or organizational invariance and multiple realization at all. Obviously, one is free to define any notion in whatever way one likes. But the classical functionalist examples of multiple realization, at least the ones that were supposed to support the autonomy of special sciences (Fodor 1974) cited phenomena that had the same capacities but different causal properties. Interestingly, Fodor (1968) did acknowledge an important distinction between two kinds of equivalence of simulations with the phenomena being simulated; weak equivalence, which is restricted to input-output relationships, and strong equivalence, which involves the equivalent causal process as well. The rub is that only strongly equivalent simulations are really explanatory of empirical phenomena. A weakly equivalent simulation only proves that it is possible to implement the capacity in some other way, but that is not the point of simulation at all. Reengineering is not about proving that some other way of bringing about a capacity is possible; reengineering is about replicating the organizational structure in a new form.

There is an important similarity between the relationship that holds among instances of the same logical structure in IBM 709(0) computers and cognitive simulations. If cognitive reengineering succeeds, then a cognitive simulation will actually *have* the capacity, not merely describe it. If the simulation is strongly equivalent, then the capacity will be present in virtue of the same (or very similar) causal structure; if it is only weakly equivalent, then the capacity might be produced in some other way–but using the same input data, it will yield the same output data as the strongly equivalent simulation. If you think of computation of IBM machines in a mechanistic way, namely in terms of levels of constitution (Craver 2007), then you might talk of equivalences at different levels of organization of a mechanism. The two computers are strongly equivalent at the computational level but not at the constitutive, electronic level of organization.

A capacity that is not substrate-neural cannot be *simulated* by building its replica in another medium at all. Reengineering makes no sense in such a case, as you cannot instantiate the capacity in a *new form* of another kind. For example, being-made-of-Swiss-cheese is not a substrate-neutral property, even if there are multiple kinds of Swiss cheese. You cannot make Swiss cheese out of apples or transistors.

All information-processing relies on such organizationally invariant properties. Whenever information-processing is causally relevant for

functioning of a physical system, the system may be fruitfully simulated. This is not to be read as saying that cognitive science may always use simulation for all cognitive capacities; I am not saying that information-processing is all there is to cognition. In particular, the physical properties of sensory apparatus are less organizationally invariant than the information-processing properties, and that may limit the scope of the possible physical realizations of the apparatus. It may turn out to be the case that only a single physical way of realizing some sensing process is viable physically or technologically, even if it is logically possible to realize it in many ways.

To summarize this part of the discussion, both computer and robotic models of cognition rely on its substrate-neutrality. Simulation makes sense only for capacities that can be instantiated using the same causal topology in different physical ways, especially if it can be simplified when instantiated (to make the simulation more understandable than the *simulandum*).

Let's now turn to computer and robotic simulations in cognitive science.

3. Simulation as Cognitive Reengineering

It is not at all controversial to say that cognitive simulation is used in cognitive science. Herbert Simon and Allen Newell (1958) even went so far as to predict that in ten years, most psychological theories will be presented as computer programs, and some ten years later, when one looks at methodological papers, computer simulation is indeed classified as a standard tool in this field (Frijda 1967, Fodor 1968). While it would be certainly hard to defend the view that in 1960s most papers in psychology were presented as computer programs or as statements about programs, as experimental psychology or personality theories remained unaffected, there was a considerable body of substantial research that followed this path.

Similarly, that computer simulation is a kind of reengineering hardly needs any special justification. With this research methodology, cognitive capacity is reverse engineered, or decomposed into its component parts. For example, Newell and Simon (1972) decomposed human problem solving into individual operations that corresponded to statements of subjects in their verbal reports (and to their eye movements). Then, the operations were analyzed as a sequence of steps included in the search for the solution in the problem space, and replicated correspondingly as a computer program. The performance of the computer program was then

empirically validated by comparing it with verbal reports from the human subjects or with eye tracking data.

One could argue that there is not so much gain in understanding models in cognitive science in terms of reengineering, as we already know that these models are complex, that they represent capacities, and that they are idealizations rather than mere abstractions. However, my point regarding the notions of "reverse engineering" and "reengineering" is not that the definitions themselves are informative. It is the practice that can be used to discover heuristics, or even normative principles if we are lucky, for simulation. By focusing on actual simulation in cognitive science and on reengineering, we can bring forth some of its criteria for the adequacy of modelling success, which will be a step forward to understanding epistemology of simulation as such.

Understanding the goal of simulation as *reengineering*, or replication of cognitive capacity in a new form, has a philosophically important consequence. Replication of the capacity guarantees that the model is really complete, which is required by the norms of mechanistic explanation. Incomplete representations of mechanisms, called "mechanism sketches" are not satisfactory (Craver 2007) as they may ignore causal factors that are relevant for the functioning of the mechanism. The only way to make sure that we understand a mechanism and have its complete causal model is to replicate the mechanism in a different medium. Note: I am *not* claiming that understanding of the mechanism is guaranteed by reengineering it. But it helps to see whether the model is complete or not. As Dretske (1994) once said, if you can't make it, you don't know how it works: this is just a negative test. (Obviously, there might be technological problems with making something that we understand but we could still know why we cannot build it anyway; for example, current technologies do not permit modelling biological organisms using the models of the same scale as original biological entities.)

The existence of the working simulation is also proof of the completeness of the mechanism, even if the mechanism is simulated only as a rough approximation. How it is possible to have complete models and to make them incrementally more precise is the topic of the next section.

4. Robotic Reengineering

Computer programs are not the only way to reengineer cognitive capacities, however. An alternative that is also interesting from the mechanistic point of view is to use robots to simulate animal behaviour (Webb 2001, 2008). In particular, these robots might be physically instantiated, not just simulated as virtual entities *in silico* (to use the simulationist jargon), which makes them physical models, just like wind tunnels whose purpose is to explore aerodynamic properties.

The distinction between virtual entities and robots can be understood as a difference between representational and immediate simulations. The representational simulations are the ones where a complex representation of a phenomenon is created, e.g. a digital simulation in a computer. There are only a finite number of features being represented: a computer simulation of weather, for example, does not represent all the physical features of rain, and those features cannot be found in the simulation (see Krohs 2008). Immediate simulations are used to directly model the phenomena using some physical resources, but it is not to say that all the physical properties of the simulator are relevant for the simulated phenomenon; the colour of the paint on the outer part of the wind tunnel is irrelevant, for example. Only some physical properties are crucial; others are not. Also, the immediate simulation, for technological reasons, is usually of limited resolution, as our measurements and technological manipulations are of limited precision.

Note that it may be hard to decide empirically what kind of simulation we deal with: immediate or representational; it's because immediate simulations are also representational, so it's not a simple dichotomy. Also, one may treat Newell and Simon's simulation of human problem solving as reengineering, which implies that it's immediate, but a weaker interpretation is of course also admissible. Of course, Newell and Simon, as defenders of artificial intelligence, intended their simulations to be immediate; their systems were supposed to think just like humans. But intentions of the researchers notwithstanding, one could still doubt whether their simulations are not only representational.

Some robots aren't even representational. Not all robotic models in cognitive science serve the purpose of explaining empirical targets that they represent: for example, *animats* are supposed to be models of possible imaginary creatures. For some, this makes them harder to evaluate (Webb, 2009); other localize them in a different place in the modelling ecosystem (Barandiaran & Chemero 2009). More importantly, some of these animat models might not be intended as explanatory at all, so they are not instance

of reengineering at all. I leave such models for another occasion; note that they might be rather instances of *forward* engineering in cognitive theory.

Let's return to reengineering. Both for behavioral and biological sciences, robotics offers a way to explain the capacities of a mechanism by building robotic models. As in other cases of simulation research, they are representations of the phenomena that are under study. In addition to their representational role, however, they are immediate simulations. This is possible only because they share the relevant relational structure with what they represent. In other words, the simulated phenomenon must be organizationally invariant.

Robots are especially useful where purely computational models are not sufficient. This can be vividly illustrated with the explanation of phonotaxis in crickets (Webb 2008). Barbara Webb and her collaborators built a robotic simulation of a female cricket that is sensitive to male chirps and moves accordingly to the auditory information it receives. The crucial part of the simulation was a physical replica of cricket ears: the ears of this insect are especially well-designed for the task of mate-finding. Namely, they have four eardrums, one pair located on the fore knees, and the other at the back of the cricket. They are connected to a tracheal tube in a way that engineers call a "pressure-difference receiver", which makes it much easier to achieve good directionality of hearing. Were the cricket simulated only in a computational way, the researchers could have to stipulate much more computational power in the insect as it would have to process more information to achieve good directionality. However, it is the physical embodiment that makes the task easier. In other words, simulation of sensory stimulation is a special virtue of the robotic models. The neural processing is simulated computationally, just like in traditional cognitive simulations, but this is not a necessary requirement of robotic modeling:

> While a variety of new and yet to be developed technologies are needed to replicate the physical interface of animals to their environment, it is generally assumed that the internal neural processes connecting sensors to actuators can be adequately replicated with electronic computation. This may turn out not to be true. Perhaps there are explicit properties and capabilities that can only be obtained by chemically identical processes (Webb 2008, 23)

Webb's robotic simulation of a cricket is clearly a Galilean idealization (Weisberg 2007, Nowak 2000): the neural system is simplified, and the motor commands were initially sent to wheels rather than legs as that was not a critical part of the simulation, so it could have been simulated in a much simplified–in engineering terms–form. What is important is that

relevant organizational properties are sufficiently similar to the ones in the biological cricket, so that we may describe, explain and predict the capacity to move towards the source of chirps when we know the activity of the component parts of the insect. The strategy that Webb uses is incremental: she started from a fairly crude model, only to add more and more biologically faithful details in subsequent simulations. They were all complete working models but the grain of simulation was finer and finer.

The model is considered to be explanatorily satisfactory when it goes beyond existing behavioral or neural data; but to build a working simulation, one needs to perform studies that were never performed by biologists before because they were not building a faithful complete model of the mechanism. For this reason, incrementally more faithful models suggest new experiments on crickets, and new experiments lead to more faithful models. In other words, the development of the model should be considered as a cyclical activity rather than a one-shot performance. The first models are sure to fail empirical validation. But instead of throwing them away, which would be recommended by a (caricature of) Popperian methodology, it is useful to tweak the model and to further reengineer it.

The interplay between behavioral and physiological studies and biorobotics is also the answer to the worry raised by Frijda (1967): complete simulations go beyond existing knowledge, and multiple ad hoc additions are needed to make them work. By validating these additions with new experimental data, we can legitimize their role in a model as working hypotheses. Ad hoc additions are then no longer hidden kludges that make validation of the theory harder; instead they should be tested independently–and thereby stop being purely ad hoc.

As inspiring and interesting as biorobotics is, it is not a universal tool. Technological limitations of a purely engineering nature make it impossible to build complete models of complex animals. Moreover, for some uses, a biorobotic model might be less faithful than a pure computational simulation. A robotic model of rat navigation (Burgess et al. 1998) is a case in point. Rats are capable of dead reckoning, that is, they are able to return to their starting position by constantly updating their cognitive map of the environment. The way they do it relies only on the signals from the vestibular system and their own motor commands; they need no further sensory stimulation. Now, the model build by Burgess, impressive as it is, does not offer any particular advantage over faithful computational models of rat navigation, such as the one offered by Conklin and Eliasmith (2005).

Biorobotics can indeed be considered an exercise in reverse engineering and reengineering: it explains the cognitive or behavioral capacities in a

mechanistic way, and replicates the mechanisms it hypothesizes in a new form. It shows both the advantages–building complete explanations, asking new questions from the perspective of the whole system–and limitations of this approach, related mainly to what we can achieve technologically. Simply put, some things are easier to simulate on a computer than to replicate physically; some are easier to do physically. It was hard for Gaudi to compute the structure of Sagrada Familia, so he used a physical model. The wind tunnel is easier to build than to simulate; but it is easier to simulate the weather on a computer than to simulate the Earth's atmosphere physically.

5. Reengineering and Dealing with Complexity

I hope that it is now sufficiently plausible to say that biorobotics is engaged in reengineering when it builds robotic models of animals. But is there anything to be gained from adopting this perspective on model-building in biorobotics? I claimed that in this way, two philosophically relevant issues may be resolved: you can substantiate the assertion that simulation, including embodied simulation, relies on substrate-neutrality rather than on multiple realization; and building immediate computer and robotic models is a way to guarantee satisfaction of a relevant methodological norm of mechanistic explanation, namely completeness of the description of the mechanism (modulo various idealizations, as models can be built incrementally, as Webb clearly shows). These are important points; nonetheless, a researcher in biorobotics may be unimpressed. Is there anything intrinsically important to reengineering that biorobotics itself would find illuminating, new or important?

On the one hand, biorobotics seems to be quite aware of the fact that it uses current engineering methods to build robots, and no illumination on this point seems to be forthcoming from reengineering. Yet there are some general points on reverse engineering that Chikofsky and Cross (1990) make which seem to be important for building models. They list six objectives that need to be taken care of with increasing complexity of software. The list applies to models in cognitive science as well. I will go step by step.

1. *Cope with complexity*. It is quite obvious that we need to develop tools that facilitate dealing with the "sheer volume and complexity of systems". Developing auxiliary tools to analyze architectures of biological systems and build robots by matching ready-made designs with anatomic parts might be an example.

2. *Generate alternate views*. It is important to create different representations of the simulated system; these representations need not, in contrast to the resulting model, be complete. This practice is legitimized in multiple models idealization as advocated by Levins (1966). Building multiple views is also recommended as a way to deal with confirmation bias, or the psychological tendency to ignore evidence that does not support one's hypotheses (see Farell & Lewandowsky 2010).

3. *Recover lost information*. Chikofsky and Cross point out that documentation of software systems usually becomes outdated in the long run. This is true also of all simulation efforts themselves, by the way; yet the analogy here is with evolution. The products of biological evolution tend to be very complex and their complexity cannot be directly related to adaptive pressures of environments. Reverse engineering helps to recover the information about possible environments where functioning of animals was adaptive. This is not necessarily linked with any optimality assumptions at all; we may as well presuppose that evolution merely satisfices, to use Simon's term (for a defense of the satisficing view of reverse engineering, see Gilman (1996)).

4. *Detect side effects*. "Both haphazard initial design and successive modifications can lead to unintended ramifications and side effects that impede a system's performance in subtle ways" (Chikofsky & Cross 1990, 16). In other words, we may discover true invariant generalizations about good designs by detecting certain side effects; this way we would know what is constitutive of cognitive capacities, and what simply co-occurs with them.

5. *Synthesize higher abstractions*. Developing generalizations at a highly abstract level is important both for engineering and theory; ultimately, we build models not only for their own sake but to discover certain general principles of cognition that apply to the broadest class of cognitive systems possible while remaining informative at the same time.

6. *Facilitate reuse*. Reverse engineering in computer science may facilitate reuse of old software; in biorobotics and simulation, it may facilitate reuse of ideas in modeling. Development of public repositories of software models and standard physical baseline frameworks (they may be as simple as LEGO Mindstorms) is a step towards replicability of results. Without it, reports about experiments on robots may remain anecdotal evidence.

6. Conclusions

In this paper, I argued that there is a close affinity between reengineering and simulation techniques used in cognitive science, and because of this, the same theoretical framework – or at least metaphor – of "reverse engineering" may be applied fruitfully to cognitive research. In particular, embodied biorobotics is a vivid example of the analogies between the two disciplines. There is a unifying framework of methods of discovery that is used by both of them.

By focusing on the similarity between reengineering and simulation, I showed that they both rely on the presupposition that what you can engineer is what is organizationally invariant. Moreover, both lead to the development of working models of systems, and working models satisfy the requirement of completeness of mechanistic models. In particular, when one adopts the incremental methodology advocated by Webb, it is possible to empirically validate the additional assumptions, needed for the models to work. They are no longer problematic ad hoc additions but preliminary empirical hypotheses to be tested in due time.

I also argued that the objectives of reverse engineering as such may be used to guide simulation research as well. They may be starting points for developing philosophical accounts of simulation science, especially of the practices of long-term validation and the relationships between the theory and simulations. I agree with Winsberg (2010) that there are important lessons for philosophy of science from computer (and robotic!) simulation.

A reengineering view on cognitive simulation is not quite so revolutionary, however. It seems to be consistent with the mainstream views on explanation in cognitive science, even if it goes beyond them. Let me elaborate.

Reverse engineering as the practice of decomposing a system in order to gain understanding about its function has been philosophically vindicated by functionalism in 1960s, in various versions and under different labels, be it functional analysis (Cummins 1975), homuncular functionalism (Lycan 1987, Dennett 1987), or neo-mechanism (Bechtel, 2008). The early flavours claimed that computer programs provide explanations of how it is possible to have a cognitive capacity (for example, Newell & Simon 1972, Cummins 1983), and in time, the focus moved to explaining how cognitive systems actually operate (Craver 2007, Bechtel 2008).

The crucial distinction between the classical functionalism and the neo-mechanism can be spelled out in terms of the difference between weak and strong equivalence (Fodor 1968): the explanatory value of the latter can be vindicated only in a mechanistic framework (for more detail, see

Miłkowski, forthcoming, chapter 3, Piccinini & Craver 2011). The perspective of reengineering fits naturally into the latter, and cannot be understood in full according to the first: namely, under classical functionalism, all that is important is that the systems are functionally isomorphic but functional isomorphism is usually so liberal that any kind of decomposition is fine. This is why Cummins talks of functional *analysis*: the hypothesis of the functional structure is a result of logical or formal analysis only, and it is not corroborated (and not verifiable) empirically at all. But researchers such as Webb or Burgess care about empirical evidence about the structure of the animals they investigate. They rely on experimental results concerning the structure of the neural system, for example; and they take pains to simulate it in a biologically faithful way. Reverse engineering as such recommends this strategy rather than engineering a new system that has the same capacity as the extant one. In computer science, reverse engineering is not about seeing how it is possible to create algorithms that perform some function but about understanding how existing structures implement some algorithms.

At the same time, the reverse engineering approach–and the mechanistic approach in general–might be seen as detrimental for the search of the general principles governing cognitive processes (Chater and Brown 2008, 38). Namely, if one is busy replicating mechanisms, one also may not see the forest for the trees. This is not an inherent danger, especially if the methodology involves the search for invariant principles–and cognitive robotics, contrary to appearances, might be used to integrate and unify different theories of cognition by requiring decisions about the cognitive architecture to be made (D'Mello and Franklin 2011, Morse et al. 2011).

Reverse engineering – and especially reengineering – sheds light on the nature of cognitive simulations, including biorobotic ones. This topic is especially important for philosophy of cognitive science in its mechanistic version, even if these engineering perspectives offer relatively minor insights–we still have to deal with problematic relationships between the theory and the simulation, problems of empirical verification and theory-ladeness of simulations: for example, multiple simulation models can be built based on the same general theory, and they may be mutually inconsistent because of necessary additions; if one fails, but contains empirically valid additions (not ad hoc ones), it does not need to imply that the original theory was incorrect. Using these insights, however, we can hope for incremental progress in the philosophical account of explanation in cognitive science.

References

Barandiaran, X. E., and A. Chemero. 2009. "Animats in the Modeling Ecosystem." *Adaptive Behavior* 17 (4): 287-292.

Bechtel, William (2008). *Mental Mechanisms*. New York: Routledge.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. *Discovery*. Princeton: Princeton University Press.

Burgess, Neil, James G Donnett, and John O'Keefe. 1998. "Using a Mobile Robot to Test a Model of the Rat Hippocampus." *Connection Science* 10 (3-4): 291-300.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive science*, *32*(1), 36–67.

Chikofsky, E.J., and J.H. Cross. 1990. "Reverse engineering and design recovery: a taxonomy." *IEEE Software* 7 (1): 13-17.

Conklin, John, and Chris Eliasmith. 2005. A controlled attractor network model of path integration in the rat. *Journal of computational neuroscience* 18, no. 2: 183-203.

Clark, A. 2001. *Mindware: An introduction to the philosophy of cognitive science*. Oxford: Oxford University Press, USA

Craver, Carl F. 2007. *Explaining the Brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, *72*(20), 741-765.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, Mass.: MIT Press.

—. 1994. "Cognitive science as reverse engineering several meanings of 'Top-down' and 'Bottom-up'." *Studies in Logic and the Foundations of Mathematics* 134: 679-689.

—. 1995. *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

D'Mello, S., & Franklin, S. (2011). Computational modeling/cognitive robotics complements functional modeling/experimental psychology. *New Ideas in Psychology*, *29*(3), 217–227.

Dretske, Fred. 1994. "If You Can't Make One, You Don't Know How It Works." *Midwest Studies in Philosophy* 19 (1): 468-482.

Eilam, Eldad, and Elliot J Chikofsky. 2005. *Reversing: secrets of reverse engineering*. Indianapolis, IN: Wiley.

Farrell, S., and S. Lewandowsky. 2010. "Computational Models as Aids to Better Reasoning in Psychology." *Current Directions in Psychological Science* 19 (5): 329–335.

Fodor, Jerry A. 1968. *Psychological explanation: an introduction to the philosophy of psychology*. New York: Random House.

—. 1974. "Special sciences (or: The disunity of science as a working hypothesis)." *Synthese* 28 (2): 97-115.

Frijda, Nico H. 1967. "Problems of computer simulation." *Behavioral Science* 12 (1): 59-67.

Gilman, Daniel. 1996. "Optimization and simplicity: Computational vision and biological explanation." *Synthese* 107 (3): 293-323.

Kitcher, Philip. 1989. Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation*, ed. Philip Kitcher and Wesley C. Salmon, 505:410-505. Minneapolis: University of Minnesota Press.

Krohs, Ulrich. 2008. How Digital Computer Simulations Explain Real-World Processes. *International Studies in the Philosophy of Science* 22, no. 3: 277-292.

Levins, Richard. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54 (4): 421-431.

Lycan, William G. (1987). *Consciousness*. Cambridge, Mass.: MIT Press.

Machamer, Peter, Lindley Darden, and Carl F Craver. 2000. Thinking about Mechanisms. *Philosophy of Science* 67, no. 1: 1-25.

Miller, George 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, Marcin. Forthcoming. *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press / Bradford Books.

Morse, A. F., Herrera, C., Clowes, R., Montebelli, A., & Ziemke, T. (2011). The role of robotic modelling in cognitive science. *New Ideas in Psychology*, *29*(3), 312–324.

Newell, Allen, and Herbert A Simon. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nowak, Leszek. 2000. "The Idealizational Approach to Science: A New Survey." In *Idealization X: Richness of Idealization*, ed. Leszek Nowak and Izabella Nowakowa. Amsterdam / Atlanta: Rodopi.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

Piccinini, Gualtiero, and Carl Craver. 2011. "Integrating psychology and neuroscience: functional analyses as mechanism sketches." *Synthese* 183 (3): 283-311.

Polger, Thomas W. 2004. Neural Machinery and Realization. *Philosophy of Science* 71, no. 5: 997-1006.

Richardson, Robert C. 2003. "Engineering Design and Adaptation." *Philosophy of Science* 70 (5): 1277-1288.

—. 2007. *Evolutionary Psychology as Maladapted Psychology*. Cambridge, Mass.: MIT Press.
Rosen, R. (1991). *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life*. New York: Columbia University Press.
Schierwagen, A. (2012). On reverse engineering in the cognitive and brain sciences. *Natural Computing*, *11*(1), 141–150.
Shapiro, L.A. 2000. "Multiple realizations." *The Journal of Philosophy* 97 (12): 635–654.
—. 2004. *The Mind Incarnate*. Cambridge, Mass.: MIT Press.
—. 2008. "How to Test for Multiple Realization." *Philosophy of Science* 75 (5): 514-525.
Simon, Herbert A. 1981. "Cognitive science: The newest science of the artificial." *Cognitive Science* 4 (1): 33-46.
Simon, Herbert A, and Allen Newell. 1958. "Heuristic Problem Solving: The Next Advance in Operations Research." *Operations Research* 6 (1): 1-10.
Webb, Barbara. 2001. "Can robots make good models of biological behaviour?" *The Behavioral and brain sciences* 24 (6): 1033-50; discussion 1050-94.
—. 2008. "Using robots to understand animal behavior." *Advances in the Study of Behavior*, 38:1-58.
—. 2009. "Animals Versus Animats: Or Why Not Model the Real Iguana?" *Adaptive Behavior* 17 (4: 269-286.
Weisberg, M. 2007. "Three Kinds of Idealization." *Journal of Philosophy* 104 (12): 639–659.
Wimsatt, W. C. (1997). Aggregativity: reductive heuristics for finding emergence. *Philosophy of Science*, *64*, 372–384.
Wimsatt, William. 2002. Functional Organization, Analogy, and Inference, in *Functions. New Essays in the Philosophy of Psychology and Biology*, ed. Andre Ariew, Robert Cummins, and Mark Perlman, Oxford UP: 173-221.
Winsberg, Eric. 2010. *Science In the Age of Computer Simulation*. Chicago and London: University of Chicago Press.