Regarding the Mind, Naturally

Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental

Edited by

Konrad Talmont-Kaminski and Marcin Milkowski

CAMBRIDGE
SCHOLARS
PUBLISHING

Regarding the Mind, Naturally:
Naturalist Approaches to the Sciences of the Mental,
Edited by Konrad Talmont-Kaminski and Marcin Milkowski

This book first published 2013

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-4341-5, ISBN (13): 978-1-4438-4341-6

TABLE OF CONTENTS

INTRODUCTION

NATURALIZING THE MIND

MARCIN MIŁKOWSKI
AND KONRAD TALMONT-KAMINSKI

> The philosophical category of category mistakes is a mistake about categories.
> —Paul Thagard (2009)

1. Kinds of Naturalized Philosophy

It has become commonplace to trace the beginnings of contemporary naturalism in philosophy to Quine's essay on naturalized epistemology (1969). As in all clichés, there is some truth to it, but reality is much more complicated. One could trace not one but two kinds of naturalism in contemporary philosophy to Quine. And, what is more, there were philosophers who practiced naturalized philosophy much earlier than Quine. It would be apt to say, therefore, that naturalists returned to the mainstream when the proponents of the linguistic turn found themselves in a cul-de-sac (Kitcher 1992).

Even so, returning to Quine's manifesto is still very useful. It allows us to disentangle the two kinds of naturalism that sprung from it which makes it possible to clarify which kind of naturalistic approach is taken by the authors of the chapters included in this volume.

The first kind of contemporary naturalism is interested with reducing all knowledge to something fundamental, be it fundamental physics or sensory stimuli. Such was the attitude of logical positivism whose heritage is quite clearly visible in Quine's thinking. In particular, when Quine proposed naturalization of epistemology, he did so because of the miserable failure of the effort to logically derivate all theories from sensory experience. Instead of showing how all theories follow logically from sensory stimuli (or rather from sentences associated by reinforcement with certain sensory stimuli), we should look at the causal processes by which theoretical knowledge is built from sensory stimuli.

Even if this kind of program may still sound plausible to many philosophers today, it is not naturalistic in the second sense of the word explored here. Namely, it takes for granted that knowledge is indeed based on sensory observation, and that epistemology should be busy with showing how sensory stimulation becomes theoretical knowledge. Unfortunately for Quine, however, this is just a dogmatic assumption. As Fodor forcefully argued, experimental science does not require that empirical evidence stem from sensory stimuli at all (Fodor 1991). We may easily replace a human being, whose sensory abilities are quite limited, with a machine taking measurements, and the experimental evidence is no less empirical for that. You don't need qualia or whatnot to make evidence more observational than it already is when measured by an automated system. Fodor, in developing a program within this second kind of naturalism, simply finds no place for dogmatic empiricist assumptions when theorizing about knowledge. Instead of making such assumptions about empirical knowledge, we should look at how science really proceeds.

Notably, however, even the kind of naturalized epistemology that did rely upon these dogmatic presuppositions went further than many of today's 'naturalist' philosophers would ever want to go. This is because there is yet another kind of naturalism – should we say a 'deflationary' one? – that conflates naturalism with ontological physicalism. Instead of showing how knowledge (or some other philosophically relevant theoretical entity) is brought about by processes that are empirically investigable, it is busy with creating theoretical frameworks to describe the relationships between such entities and 'the physical', where 'the physical' is usually either left almost completely without any content, or equated with the view that the fundamental cement of the universe are some physical entities. These entities are usually couched in a terminology that suggests that they be elementary and atomic rather than relational. (That the latter view, presupposed for example by Jaegwon Kim in most of his writings, is hardly part of contemporary physics, barely needs mention; cf. Ladyman, Ross *et al*. 2007.)

So, to sum up, we may distinguish three kinds of naturalism: (1) the one that uses science to argue for philosophical positions even if they are clearly at odds with what science says or does; (2) the one that sees science and philosophy (along with the humanities) as belonging on the same continuum, and that does not seek for strict boundaries between those; and (3) the one that is just a new label for physicalism.

Since it is the second kind of naturalism that is of most relevance to the authors included here, it makes sense to elucidate it a bit further. It should be clear that this naturalism is a methodological position rather than defending any particular ontological claim. It is also not focused on purely philosophical questions, such as showing how physicalism or empiricism is true by citing scientific papers and textbooks, or performing spectacular thought experiments about a future, ideal and complete fundamental physics. Rather, it is interested in the real subject matter of science, so it takes science most seriously in allowing that scientific discoveries lead to conceptual revolutions. Notably, it may also point to conceptual conundrums in science that seem to stem from philosophical assumptions or address worries connected with philosophical problems that seem pertinent to scientific theories, such as the worry whether causal explanations are genuine in a certain domain or whether some entities, say mental, are causally efficacious for certain phenomena. But these worries and conundrums are of common interest for philosophers and scientists alike. In effect, this brand of naturalism does not posit any abrupt discontinuity between scientific knowledge and philosophy.

In philosophy of science, this kind of approach is attributed to Kuhn (1970) and pioneered by Fleck (1979), who insisted on investigating the practice of science instead of proposing rational reconstructions of the logic of inquiry. Most philosophers of science have followed suit (for a pluralistic view on naturalized philosophy of science, see Callebaut *et al.* 1993.) Some have developed Kuhn's (1970) historical methods while others have turned to a cognitive approach to understanding science (e.g., Giere 1994, Nersessian 2008). Both the historical and cognitive approaches agree that naturalistic philosophy of science should not restrict itself to looking at the products of science–journal papers, books, reports, models or whitepapers–but must consider science as a process. It is not only important to consider whether theories are properly justified but also to think about how those theories come into being. By looking at how discoveries are made, we may improve our understanding of such important issues as how the research heuristics of localization and decomposition make possible certain kinds of reduction (Bechtel and Richardson, 1993). Or we may simply help improve research strategies.

Research into discovery in science belongs to the tradition initiated by Herbert A. Simon and his collaborators (see, for example, Langley *et al.*, 1987), and its influence on many subsequent philosophers of cognitive science and neuroscience is beyond doubt. In particular, the mechanistic philosophy of science (Darden, 2006; Craver, 2007; Bechtel, 2008) follows in Simon's footsteps by investigating the scientific heuristics employed in the identification of mechanisms; but you don't have to focus solely on mechanisms to talk of the importance of heuristic methods (Wimsatt, 2007). At the same time, this kind of philosophy of science is deeply entrenched in philosophy of biology, where evolutionary theory supplies the conceptual framework.

This volume is intended to be an exercise in this kind of naturalism, a naturalistic philosophy that does not need to be naturalized further because it is methodologically naturalist. In particular, in theorizing about the mind, the discussion is not focussed on traditional topics of the philosophy of mind, such as finding aprioristic arguments showing that it may be logically possible to reduce future psychology to future complete physics. Fascinating as such arguments are, they are often of secondary interest to people who want to know something about how minds work, what they are, and how best to investigate them. We do not wish to suggest that all philosophy of mind is futile scholasticism to be replaced with science, so long as there remains a distinctive role for philosophers of psychology and cognitive science to play. This statement calls for some elucidation, so it makes sense to see what would be the point of such a wholesale, radical rejection of philosophy of mind.

> The philosophy of mind is over. The two main debates in the philosophy of mind over the last few decades about the essence of mental states (are they physical, functional, phenomenal, etc.) and over mental content have run their course. Positions have hardened; objections are repeated; theoretical filigrees are attached. These relatively armchair discussions are being replaced by empirically oriented debates in philosophy of the cognitive and neural sciences (Chemero and Silberstein, 2008, p. 1).

But Chemero and Silberstein are clearly wrong. For example, the debate over mental content is not over at all, but is still as fierce as ever (Ramsey, 2007). Admittedly, for Chemero, the arch-antirepresentationalist, this debate is best swept under the carpet but it is a gross simplification (if not distortion) of the practice of cognitive science to say that representations do not play any role at all in cognitive explanations. They do. And we still do not understand what that role is, exactly. Similarly, the discussions over functionalism, and especially about reduction, multiple realization and causation are very much relevant both for today’s philosophy of mind and philosophy of psychology (see, for example, Shapiro, 2004 or Polger, 2004). So while it might be true that re-focusing the debate on the goals of the scientific research into mental capacities should render the debate more concrete than in the past, when it relied on intricate thought-experiments, this does not make the whole field of traditional philosophy of mind irrelevant to naturalistic philosophy.

Paul Thagard (2009) gives at least two reasons why philosophy is relevant to cognitive science, and they apply to philosophy’s significance for psychology or any science, for that matter. Firstly, philosophical thinking in the context of science remains at a higher level of abstraction, or generality, which makes it easier to discover commonalities that would have been occluded by narrow specializations. This is particularly important in interdisciplinary conglomerates such as cognitive science, where merely verbal disputes are bound to arise because of the differences in terminology in various fields. Secondly, philosophers are skilled in normative arguments, and when science meets practical application–especially in the fields where practical advice is sought, such as psychiatry, neurosurgery, economics, or educational policy–philosophers are able to clear up the ways norms may be inferred. For example, instead of presupposing a traditional model of ideal rationality exemplified by homo oeconomicus, naturalistic philosophers may also link practical applications with the idea that human beings are merely satisficers.

Philosophical 'therapy', sermonized by late Wittgenstein and his followers, that aims to remove philosophical or metaphysical vestiges from science by declaring them meaningless, is itself quite meaningless for scientists. The trivial empirical observation that in some natural language words do not mean the same thing as in the technical vocabulary of science is of hardly any importance. Neuroscientists, for example, use the intentional idiom to talk about brain parts, and Wittgensteinians think that this is a serious category mistake, as only persons may be ascribed intentional capacities. The argument is that in natural language we don't say so. As this is an outright falsity: it takes little effort to see that people frequently say, for example, that one eye sees a slightly different image than the other.[1] Of course, one could interpret this kind of utterance as an abbreviated form of *the* correct "a human being sees with one eye differently than with another," and paraphrase similar utterances in the same way. Otherwise, this frequent talk would be a category mistake, as it is only the whole person that sees. However, the mere fact that one would need to paraphrase–and in the way that is so prolix as to make such paraphrase sound artificial–means that what is at stake is not ordinary language but a certain kind of regimentation of natural language into some kind of theoretically-laden talk. The language usage is simply different, and the biographical observation that (some) Wittgensteinians do not consider frequent English usage as standard, preferring hypercorrection over standard usage, does not justify the claims about category mistakes.

[1] If still in doubt, see for example *Corpus of Contemporary American English*, available online at http://corpus.byu.edu/ (Davies 2008). The query "person sees" has 53 matches, "eye sees" has 54 (you might also add "eyes see" with their 67 matches, and "persons see" with 1 match). The Google Books Historical Corpus allows investigating the changes in usage on the same website.

Wittgensteinians are free to talk whatever way pleases them but it is not a philosophical argument for others to change their language. There is nothing special in (a hypercorrect flavour of) ordinary language that would make it a privileged source of insight. The theoretically-laden paraphrase, in other words, has to be justified, and mere linguistic facts do not support it.

For a similar reason, Thagard argues that supposed category mistakes such as this are just the figments of philosophical imagination: concepts change and are not set in stone once they occur in ordinary speech. So, concepts are not something that you could analyze to gain philosophical insight into how things are (Fodor 1998); at best, you get lexical semantics. Indeed, the very category of 'concept' may not be as unified as philosophers usually presuppose (Machery 2009).

At the same time, it needs to be stressed that some of Wittgenstein's insights turned out to be useful in the scientific investigation of prototype concepts. In particular, the notion of family resemblance was put to use there (Rosch & Mervis 1975). And, maybe ironically in this context, the Wittgensteinian insistence that the task for philosophers is to clear up conceptual confusion is still valid; but, this time, it is to clear up the mess that the Wittgensteinians themselves introduced by sermonizing that common sense is the measure of all things. For this reason, naturalists often engage in polemics with non-naturalized philosophy. We will see examples of that in this volume as well.

All that we have said up to this moment is not to be meant to imply that naturalistic philosophy of the kind that is pursued here is supposed only to serve the needs of science. There are distinct philosophical interests that naturalists take in their reflection that crosses the boundaries between the scientific and manifest image of the world. For example, for Daniel Dennett (2009) this is one of the tasks for philosophy of cognitive science: to understand the impact of new findings on the self-image of human beings, and help draw a broad, rational picture of the world. This, however, does not mean that the manifest image be replaced with some particular special science; we do not need scientific standards in everyday life, and for this reason we will satisfice by taking into account only the most crucial insights that science provides.

2. From Naturalized Philosophy to Naturalistic Thinking

There is a second reason why we opened this introduction with a quote from Quine. Some of the early versions of the papers in this volume were presented during workshops in Kazimierz Dolny, Poland that we have organized over a number of years, and a certain kind of dualism that seems to correspond to the two kinds of naturalism discussed above is reflected in the names of these workshops. Just like Quine, they started out as the Kazimierz Naturalized Epistemology Workshop (KNEW) back in 2005. After some time, roughly at the point when we decided that there was enough material about normativity to think of editing a volume about it (which appeared as Miłkowski and Talmont-Kaminski, 2010), we retained only the acronym, as we felt that epistemology was already successfully naturalized. The unofficial expansion was Kazimierz Naturalized Everything Workshop, while the official one – Kazimierz Naturalist Workshop. We wanted to stress that we are no longer so much interested in meta-philosophical reflection about the status of naturalism as in the real work done.

Because many of the participants of the workshops have decided to come regularly, we believe we can say that there is something that brings them together; this is exactly the second kind of naturalism, as described above. For the present volume, we asked some of our regulars to contribute chapters related to naturalistic approaches to the mind.

In the first section, the sciences of the mind are investigated as process, not as product: in other words, the authors focus on discovery rather solely on justification strategies. Marcin Miłkowski frames some of the simulation research in cognitive science and cognitive robotics in terms of "reverse engineering" and shows what light it sheds on the practice of cognitive investigation. The heuristics used to reverse-engineer a piece of software correspond quite strongly to the best practice in cognitive modelling. Samuli Pöyhönen, in turn, focuses on how the mental mechanisms of such complex phenomena as bulimia nervosa, which are partially socially constructed, could be still understood as natural kinds. He concludes that the notion of the natural kind, if used to discuss explanatory practices, should be relaxed as to include cultural factors as well. Finally, Mark Alfino shows that investigating the mechanisms that underlie the phenomenon traditionally called wisdom may be beneficial not only for psychology but for philosophy as well. This is because, if we are interested in improving our epistemological practices, we better look at what wise people do and how their expertise is explained by current psychological research.

In the next section, Biological Cognition, three chapters were included. Alvaro Moreno argues for understanding cognition as a biological phenomenon, and in particular one that occurs in autonomous systems capable of recursive self-maintenance. Cognition considered this way can no longer be reduced to abstract logical or computational capacities of the mind. Paweł Grabarczyk focuses on the difficulties with applying psychological vocabulary to non-linguistic animals in order to explain their behaviour. Indeed, he argues, by drawing on numerous examples from ethology as well as on purely philosophical considerations, there are immense difficulties with justifying, for example, the claim that animals perceive something as belonging to one ontological category rather than another. And Benoît Dubreuil presents research he has been conducting in recent years on a topic that has been neglected by cognitively inclined philosophers and social scientists: the evolution of political hierarchies in humans. Why do humans live in political systems as different as small egalitarian foraging bands and large-scale hierarchical societies? His view is that cognitive sciences are essential tools that help making explanations in the social sciences deeper, but that they only reveal their full potential when they are used in conjunction with more traditional methods of social research, as part of an integrative and pluralistic approach to complex social phenomena.

The third section, Realisation, Explanation, and Reduction, focuses on another set of issues. Naturalists often see traditional reductionism as too far removed from scientific practice. Instead of arguing for autonomy of special sciences by declaring them irreducible, Markus Eronen uses the notion of robustness, as defined by Wimsatt (2007), to defend the view that entities used by special sciences are real as long as they are robust, i.e., presupposed by multiple independent theoretical frameworks. In this way, he naturalizes the traditional non-reductive physicalism without being committed to any a priori dogma. In particular, Eronen rejects the traditional argument from multiple realizability. In contrast, Panu Raatikainen argues conditionally that even when one accepts that there is multiple realizability, the argument that the mental cannot be causally efficacious is faulty. In particular, the causal exclusion argument, voiced most powerfully by Jaegwon Kim, cannot be sound if one accepts the interventionist account of causation, which has been successful in analyzing causal claims in science and dealing with traditional difficulties of theories of causation. Raatikainen defends, in effect, the view that it is possible for the mental to be causally efficacious without committing oneself to type identity or similar reductionist views. Witold Hensel, on the other hand, shows that many naturalists are careless when arguing for anti-reductionism. As it turns out, Craver's (2007) mechanism, his declarations notwithstanding, is best understood as a reductive framework, not as an anti-reductionist one. We think that these three papers show a range of different approaches to reductionism in contemporary naturalism. The debate is far from over.

Section four, Metaphysics of Mind Naturalized, contains papers that target the traditional metaphysical issue of qualia or phenomenal consciousness in the philosophy of mind. Jonathan Knowles argues for a version of naturalism in which the hard problem of the phenomenal consciousness does not arise, as what the organism cognizes, however basically, is *always already* a world *for it*, filled in and specified in relation to categories that only make sense in relation to *its* particular subjectivity and needs. Tadeusz Ciecierski tries to pin down the traditional claim that qualia are intrinsic properties and investigate what its consequences are. As it turns out, in conjunction with the claim that there are so-called object-dependent thoughts, the traditional notion cannot be consistent unless accompanied by some form of direct realism or direct reference thesis. This is a quite surprising corollary of a widespread view. Dimitris Platchias, in turn, argues that the challenge posed by David Chalmers (1996) against non-dualists may be answered in terms of the HOT theory of consciousness.

The last section is devoted to naturalization of truth and correspondence. Jaime Gomez presents a vindication of a naturalised theory of concepts, linked with an isomorphism-based notion of correspondence. In contrast to traditional accounts of correspondence, he acknowledges that a role for model-like representations in autonomous cognitive systems has to be shown. Krystyna Bielecka analyzes deflationary strategies as implying certain supervenience claims to show that they are either uninformative or implicitly inflationary. Even if she focuses only on Hartry Field, this argument seems to be generalizable to linguistic deflationism as such. Maria Frappoli, focuses on the enterprise of naturalizing truth, especially on the so-called prosentential theories of truth. For this reason, she gives more naturalistic ground to positions close to deflationism (if not deflationist per se).

As is, we hope, clear, the volume shows a lively debate between different positions in contemporary naturalism. Eronen and Hensel disagree about the role of reductionism, Frappoli argues for an essentially linguistic (prosentential) account of truth, while Bielecka argues against it; Gomez defends an account of representation as isomorphic encoding which is rejected by Moreno. A convergence of opinion is to be found in dead alleys in science; so long as there are insurmountable differences and controversies, there is a chance for cognitive progress. We only hope that the volume is a step in this direction.

References

Bechtel, William. 2008. *Mental Mechanisms*. New York: Routledge, part of the Taylor & Francis Group.

—. 2009. "Constructing a Philosophy of Science of Cognitive Science." *Topics in Cognitive Science* 1 (3) (July): 548-569.

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

Callebaut, Werner G, *et al*. 1993. *Taking the naturalistic turn or how real philosophy of science is done*. Chicago (Ill.); London: University of Chicago Press.

Chalmers, David John. 1996. *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press.

Chemero, Anthony, and Michael Silberstein. 2008. "After the Philosophy of Mind: Replacing Scholasticism with Science." *Philosophy of Science* 75 (1) (January): 1-27.

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

Darden, Lindley. 2006. *Reasoning in biological discoveries: essays on mechanisms, interfield relations, and anomaly resolution*. Cambridge; New York: Cambridge University Press.

Davies, Mark. (2008-) *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/

Dennett, Daniel C. 2009. "The Part of Cognitive Science That Is Philosophy." *Topics in Cognitive Science* 1 (2): 231-236.

Fleck, Ludwik. 1979. *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press.

Fodor, Jerry A. 1991. "The Dogma that Didn't Bark (A Fragment of a Naturalized Epistemology)." *Mind* 100 (2): 201-220.

Fodor, Jerry. 1998. *Concepts. Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Giere, R.N. 1994. "The cognitive structure of scientific theories." *Philosophy of Science 61*(2): 276–296.

Kitcher, P. 1992. "The naturalists return." *The Philosophical Review* 101 (1): 53–114.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Ladyman, James, Don Ross, David Spurrett, and John Collier. 2007. *Every Thing Must Go. Metaphysics Naturalized*. Oxford: Oxford University Press.

Langley, Pat, Herbert A Simon, Gary L. Bradshaw, and Jan M. Żytkow. 1987. *Scientific discovery: Computational explorations of the creative processes*. Cambridge, Mass.: MIT Press.

Machery, Edouard. 2009. *Doing without concepts*. Oxford: Oxford University Press, USA.

Milkowski, Marcin, and Konrad. Talmont-Kaminski (ed.). 2010. *Beyond description: naturalism and normativity*. London: College Publications.

Nersessian, Nancy. 2008. *Creating scientific concepts*. Cambridge: A Bradford Book.

Quine, W V O. 1969. Naturalized epistemology. In *Ontological relativity and other essays*, 69-90.

Polger, Thomas W. 2004. *Natural Minds*. Cambridge, Mass.: MIT Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family resemblances: Studies in the internal structure of categories." *Cognitive Psychology* 7 (4): 573-605.

Shapiro, Lawrence A. 2004. *The Mind Incarnate*. Cambridge, Mass.: MIT Press.

Thagard, Paul. 2009. "Why Cognitive Science Needs Philosophy and Vice Versa." *Topics in Cognitive Science* 1 (2): 237-254.

Wimsatt, William C. 2007. *Re-engineering philosophy for limited beings: piecewise approximations to reality*. Cambridge, Mass.: Harvard University Press.

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.