Simulation and the Sense of Understanding

Jaakko Kuorikoski

Jaakko.kuorikoski@helsinki.fi

Philosophy of Science Group

University of Helsinki

1. Introduction

Computer simulation is widely taken to be the best, and sometimes the only, tool with which to study highly complex phenomena. However, in many fields, whether simulation models provide the right kind of understanding comparable to that of analytic models has been and remains a contentious issue (cf. Galison 1996; Lehtinen and Kuorikoski 2007; Lenhard 2006). Simulation models are often themselves difficult to understand and it is often noted that replacing an unintelligible phenomenon with an unintelligible model is not epistemic progress. One of the principle aims of science is the creation of scientific understanding and the ability to create understanding should be one criterion by which computer simulation techniques should be assessed. The aim of this paper is to point out that such assessment may often be hampered by a conflation between the sense of understanding and understanding proper. This confusion can distort the appraisal of simulation models in both ways. On the one hand, simulations can provide understanding without the corresponding sense of understanding. On the other hand, evocative visuals and vague intuitions about maker's knowledge can bring about a sense of understanding without actual increase in true understanding and thus result in an illusion of understanding. In order to improve our understanding of the merits and drawbacks of simulation techniques, we need to replace appeals to the sense of understanding (and vague intuitions about intelligibility that may depend on it) with explicit criteria of explanatory relevance and rethink the proper way of conceptualizing the role of a single human mind in the collective understanding of the scientific community.

Recent philosophical interest in the central role of models in scientific practice has highlighted the fact that creating understanding is not just a matter of providing more information about the phenomena to be understood. Cognitive limitations of humans set conditions on what can be understood, conditions that have to be met by using idealizations, abstractions and outright falsehoods (Teller 2001). However, the exponential increase in the computational power available to scientists and the spread of standardized simulation packages has meant that computer simulation has become an integral modeling tool in almost all sciences. In principle, simulation models are not subject to the constraint of analytic tractability and hence can do away with many of the distorting idealizations and tractability assumptions of analytic models. However, the very fact that a simulation model can do away with such idealizations often makes the simulation model itself *epistemically opaque*: the relationship between the stipulated initial conditions and the simulation result cannot be modularly decomposed into suboperations of sufficient simplicity, something that could be "grasped" by a cognitively unaided human being (Humphreys 2004, 147-150). If a computer simulation itself becomes so complex that we understand it no better than the phenomenon being simulated - when the simulation reaches the complexity barrier, as Johannes Lenhard puts it (2006) - has our understanding about the phenomenon itself increased?

The structure of the paper is the following. The following section distinguishes between the sense of understanding and understanding proper and gives a broadly Wittgensteinian deflationist account of the latter. The third section discusses the possible illusions of understanding, sensations of increased understanding without accompanying increase in understanding proper, that can ne expected to be present in the context of simulation models. The fourth section discusses two possible ways of responding to the epistemic opacity of simulation models: the use of metamodels and the reconceptualization of the boundaries of the cognitive subject. All hypothetical simulation models discussed are presumed to be “true” or giving an appropriately accurate picture of the simulated phenomenon throughout the paper; the question asked is whether a simulation can provide understanding, not whether the simulation is correct or not or how we might come to know this.

2. The concept of understanding

Many simulation models are so complex that they go beyond the limits of human understanding. The limits of understanding cannot be adequately charted until there is understanding of the concept of understanding itself. Yet understanding has thus far received little attention in the literature of the philosophy of science. Although the supposed means of conveying understanding – explanation - has for a long time constituted one of the most productive branches of the general philosophy of science, the product itself has been seen as something that cannot or should not be philosophically explicated (Newton-Smith 2000). Even the basic ontological category of understanding seems to be unclear; should understanding be conceived as a privileged mental state, as some kind of super-knowledge, as a cognitive act or even as a special method?

Although there has been considerable discussion of the concept of explanation, the supposed correlate of explanation – understanding - has been to a large extent left out of the picture (Newton-Smith 2000, 131). This curious situation may partly be the result of Hempel’s belief that the concept of understanding was pragmatic to the bone, a psychological by-product with no epistemic relevance, and that it therefore did not deserve any meticulous philosophical analysis (Hempel 1965, 413). No doubt another reason has been the dialectics of the theory of explanation, which has largely been based on repeated toy examples, historical cases and appeals to intuitions. This methodology has left the theory of explanation largely stipulative in the sense that it has ignored the question of why we want explanations in the first place; explanation just *is* the laying out of the causal history of an event (Salmon 1984) or unification of our overall world-view (Kitcher 1989).

Understanding has sometimes been associated specifically with the interpretation of meanings or intentional action and has even been seen as somehow distinct from or the opposite of explanation. My interest here is in understanding conceived more broadly, as the general objective of (the conveying of) explanations. Most strands of thinking about understanding as distinct from explanation share an aspect of understanding of understanding that is discarded here; understanding itself is thought of first and foremost as a phenomenon hidden inside the mind. This mentalist conception of understanding is also presupposed in the recent attack on philosophical theories of explanation by J. D. Trout (2002). Trout accuses theories of explanation as relying on a criterion of goodness of explanation, according to which a good explanation should produce a feeling or a *sense of understanding*. According to Trout, this sense of understanding has little to do with epistemic progress proper and is usually just a result of well-known psychological retrospective and overconfidence biases. Theories relying on such unreliable and contingent factors cannot be epistemologically interesting.

Michael Scriven already pointed out that it is absurd to identify the sense of understanding with understanding itself, since the former can so easily be mistaken (Scriven 1962, 225). Perhaps understanding should then be seen as a psychological state or activity that is not transparent to introspection of the individual in question? However, Wittgenstein persuasively argued that the grammar of understanding is not that of a state-concept in the first place, but similar to that of *an ability*: understanding is attributed according to whether somebody can reliably *do* something. Understanding is not about just possessing knowledge, but about what one can do with the knowledge and how reliably. Note that understanding is only *akin* to an ability in the sense that it would be equally wrong to think of it as a distinct species of (psychological) ability or skill (such as a mental faculty), something with a deeper underlying essence. Instead, understanding is a regulative or normative concept in the weak sense that it concerns whether people have the ability to do things *correctly* or in a *right* way. As such, understanding can only be attributed by the relevant community according to public criteria (Wittgenstein 1953 [1997], §§ 143-159, 179-184, 321-324; Baker & Hacker 2005, 357-385; Ylikoski forthcoming; Ylikoski and Kuorikoski forthcoming; see also de Regt 2004, 100-102 and Elgin 2007, 39). These criteria do not (indeed cannot) primarily concern private cognitive processes, brain states or, even less, subjective feelings, but manifest performances. Understanding is akin to an ability, not to a hidden state. It is fundamentally public, not private.

Cognitive processes (comprehension) taking place in the privacy of individual minds are a *causal* prerequisite for possible fulfillment of these criteria, but the processes themselves are not conceptually primary. They are not the criteria of understanding in the sense that we would have to know them in order to say whether somebody *really* understands something in the same sense that we do not need to know the cognitive and neural processes that enable one to ride a bike in order to judge whether one can ride a bike. In a sense the Wittgensteinian account is deflationist, since it denies that there is a deeper essence of understanding behind the manifest abilities according to which understanding is attributed. In cognitive science, mental models of different varieties are the standard ways of causally explaining reasoning and comprehension and it appears that these mental representations are not propositionally structured, but instead represent dependency relations directly or “intrinsically” (cf. Waskan 2009). Yet the postulation of these models addresses a different question, i.e., how individuals can achieve what is demanded by the public standards constitutive of the concept of understanding itself[1] and the *correctness* of the internal models is judged by the external displays of understanding, not the other way around (Ylikoski forthcoming). This is the ‘natural’ cognitive relation between the behavior of the agent and the information providing the understanding that Trout correctly leaves for empirical science to investigate (Trout 2004, 203). But this does not remove the fact that understanding in itself is a categorically different kind of thing.

Whatever computational and memory limitations these cognitive processes have also limit the possible manifestations that judgments of understanding are based on. Nor should one deny the existence of feelings of illumination or ‘getting it’ associated with increased understanding. This is the sense of understanding. The phenomenological state of sense of understanding is only a fallible predictor of future performance to be judged against the public criteria. One can judge whether one really understands something only by comparing one’s explanatory or inferential performance against an external public standard. Wittgenstein’s main concern was, of course, the understanding of the meanings of words and utterances, but his general argumentative thrust can be straightforwardly generalized to apply to explanatory understanding. Thus we arrive at a characterization that is general enough not only in including the natural and the human sciences, but also in the way it links and makes intelligible the use of the same word in practical every-day matters and language.

What can be said of the criteria that one needs to fulfill in order to be eligible for possession of scientific understanding of a theory or a phenomenon? Hank de Regt and Dennis Dieks (2005) stress the manner in which the criteria for scientific understanding have changed in the history of

[1] Jonathan Waskan seems to claim that these mental models *make* phenomena intelligible (Waskan 2009, ?), which, of course, would just lead to a regress: in virtue of what do these mental models make phenomena intelligible? This mentalistic way of conceptualizing understanding also leads easily to confusing sense of understanding with understanding (cf. ibid., ?)

science. It is certainly necessary to allow certain flexibility in the criteria of understanding so as not to render all past scientific explanations unsatisfactory or downright unintelligible if they do not meet the standards of today. However, certainly understanding *per se* is such an integral part of our conceptual scheme that a complete lack of continuity in historical applications of this concept would be hard to accept. There should also be a place left for unabashedly normative use of the concept from our current perspective; we do make judgments as to whether some particular phenomenon was understood at some given time, and the grounds for these judgments should have continuity with arguments given in some other era. After all, explanatory progress in science is at least partly a matter of empirical discovery, not just of conceptual change.

When looking for a common factor in the attributions of understanding of *a theory*, De Regt and Dieks arrive at a conclusion that the attribution of understanding rests on the subject's ability to draw qualitative conclusions about hypothetical changes in the explanatory model without making exact calculations with additional inferential aids, such as pen and paper or a calculator (ibid., 151). According to them, the criterion of understanding of theories is unaided inferential performance. Yet the ultimate goal is not just to understand theories or models, but to understand the phenomena that the theories and models are about.

The central idea of inferential performance as the constitutive criterion of understanding can be further developed by linking it to James Woodward's account of scientific explanation in the following way: Woodward's theory of explanation tells us more specifically what *kinds of inferences* are constitutive of understanding, of theories as well as of phenomena that the theories are about. According to Woodward (2003), explanation consists of tracing or exhibiting functional dependency relations between variables. Explanation is thus doubly contrastive; the functional relationship links the possible values of the *explanans* to possible values of the *explanandum*. These explanatory relationships provide understanding by giving answers to what-if-things-had-been-different questions concerning the consequences of counterfactual or hypothetical changes in the values of the *explanans* variable. These answers are the basis of inferential performance constitutive of understanding. In the case of causal explanations or explanations given on the basis of a causal model, the relevant hypothetical changes to consider are interventions, ideally surgical manipulations that affect only the *explanans* variable of interest and leave the rest of the model intact (apart from the changes caused by the change in the *explanans* variable dictated by the model, of course). Thus the conception of understanding as inferential ability is fully compatible with the realist (or ontic) idea of understanding as knowledge of causes and mechanisms, as long as causal relations are understood as dependencies invariant under interventions.

Woodward's account intimately links explanatory knowledge to our capacity to function in the world as goal-directed manipulators. Whereas the epistemic conception of explanation of the D-N – model was concerned only with predictive power, Woodward's theory stresses our role not only as passive observers, but also as active agents. Understanding is the ability to make correct inferences on the basis of received knowledge and causal knowledge concerns the effects of manipulations and therefore licenses inferences about the effects of our actions.[2] Understanding thus lies not only in correctness of inference, but sometimes also in effectiveness of action based on the information to be understood. Although a mechanic may not be able to give a theoretical description about the physical laws governing the workings of a combustion engine, he does have an intimate understanding about its workings because he can effectively manipulate it – in order to fix an engine one has to know the functional roles of the parts and the effects of manipulating them. This idea probably underlies Fred Dretske's unarticulated intuition that one cannot understand how

[2] Causal reasoning should not be seen as just a conceptual tool in arriving *at* understanding (as De Regt and Dieks see it), but as an end in itself - an important *kind* of understanding. Of course, different varieties of causal reasoning are often a causal means to understanding, even to non-causal understanding, but this is a different issue.

something works unless one can build it (Dretske 1994) and Johannes Lenhard's pragmatic conception of understanding (Lenhard 2006).

Understanding thus comes in many varieties and shades, some of which might be considered more scientific than others. However, since science is fundamentally a communal activity, extra criteria of being able to provide explanations and of being able to justify the information responsible for the understanding might be appropriate for attributions of specifically scientific understanding to individuals. Since scientific knowledge is often highly systematized into hierarchic sets of principles, the ability to justify correct inferences made on the basis of a limited set of theoretical principles, i.e., *theoretical understanding,* also characterizes much, but not all, of specifically 'scientific' understanding. Thus understanding as the ability to make correct inferences also accounts for the common intuitions linking understanding to integration of isolated facts into background knowledge (as in Schurz and Lambert 1994 and in Elgin 2007) and to unification more generally (Friedman 1974; Kitcher 1989). However, whereas unificationists have to adopt a stipulative stance in claiming that understanding simply is unification (Barnes 1992), understanding as inferential ability follows as a natural consequence from a well-ordered knowledge store or from a powerful set of argument patterns.

Understanding is epistemic in the sense that correctness is usually to be understood as truth or truthlikeness; understanding is usually seen as factive (Grimm 2006; Trout 2002).[3] Understanding is also epistemic in the sense that correct inferences about alternative possibilities are crucial in finding out new things. At the same time, understanding is pragmatic in the sense that it has an intimate connection with our non-theoretical aspirations and also in the sense that it *causally* depends on individual psychological abilities. From this perspective it is not clear why de Regt and

[3] However, it is primarily the correctness of inferences that matters and that (partial) understanding can therefore be provided by models and theories that incorporate significant falsities in the form of idealizations etc. (cf. Elgin 2007).

Dieks single out the qualitative predictions concerning consequences of changes in the explanatory model, since quantitative mathematical manipulations surely are a form or a dimension of understanding as well, usually one regarded as highly scientific (although one can, of course, stipulate that intuitive understanding must be arrived at by intuitive means). In fact, the degree of externalization of inferences and the amount of explicit calculation that would still count as constitutive of understanding vary across scientific disciplines. For example, economists are notorious for insisting on proficiency in mathematical model manipulation and for especially valuing analytical solutions done with only pen and paper as inferential aids (Lehtinen & Kuorikoski 2007).

3. Simulation and the illusion of understanding

We have thus far distinguished between the psychological phenomenon of sense of understanding from the degree of understanding itself, which can in turn be operationalized (in the good old-fashioned strong sense) as the ability to make correct counterfactual inferences about the object of understanding. The sense of understanding is only an indicator of understanding proper and it is the latter notion that is epistemically and pragmatically relevant. The sense of understanding has an important metacognitive role in providing immediate cues according to which we conduct our epistemic activities and it is also an important psychological motivating factor for those activities (Ylikoski forthcoming). However, if the ability to create understanding is to be taken as an important criterion in the final *assessment* of simulation techniques, this assessment should not be based on vague intuitions about understanding, usually dependent on the sense of understanding, or on the sense of understanding directly, unless the sense of understanding turns out to be a sufficiently reliable indicator of understanding itself.

The trouble is that empirical studies seem to indicate that people systematically over-estimate their abilities to explain the workings of mechanisms and natural phenomena. As Frank Keil and his associates have empirically demonstrated, when given the possibility of comparing their explanatory performance (inferential performance) to an external standard, people tend to downgrade their assessment of their own explanatory performance and only after learning more about the test-case do they begin to raise their self-assessment score towards the initial level. This effect is not present in the self-assessment of the subject's ability to recollect facts and hence is not simply an aspect of the alleged general overconfidence bias. (Keil 2003; Mills and Keil 2004; Rozenblit and Keil 2002) People are systematically overconfident specifically about their explanatory understanding. To the extent that self-assessment of understanding is normally based on the sense of understanding (which I see no reason to doubt), this suggests that the sense of understanding is an unreliable indicator of understanding.

The miscalibration of the sense of understanding is a general phenomenon. However, there is a possibility that the sense of understanding is especially biased in some specific contexts. In principle, the sense of understanding can be misleading in both ways: something can be understood without an accompanying sense of understanding (tacit learning) and there can be a sense of understanding without a corresponding increase in inferential ability. I will here focus on the latter possibility, the danger of *illusion of understanding* (Ylikoski forthcoming), since there are a number of reasons to expect[4] that this danger might be especially severe in the context of simulation studies.

3.1. Mistaking understanding of sub-operations for understanding of the process

[4] It is important to admit that the following worries are in the end empirical psychological hypotheses.

Even if the sense of understanding were a reliable indicator of inferential performance when applied to simple inferential tasks, the danger of miscalibration can be expected to increase when the task becomes more complex. First, simply the fact that the overall cognitive load increases probably makes self-evaluation more unreliable. Second, complex inferential tasks are something that the agent rarely encounters, thus there is less feedback and external benchmarks against which the sense of understanding could have been calibrated. In the case of simulated complex systems, both of these worries are present. An additional worry is that the sense of understanding of a simple sub-task can be mistaken as an indicator for true understanding of the more complicated task.

As Roman Frigg and Julian Reiss point out (2009), at the bottom level of the foundational transition rules on which the simulation is built on, simulations are always in principle understandable, since the changes of the states of basic elements (for example state-transition rules for single cells in a cellular automata model) follow well defined rules laid down in the programming of the simulation.[5] These rules themselves are usually simple and intuitive and thus relatively easy to understand, but the inferences concerning the results of interactions of these elements are well beyond our limited cognitive powers. The danger of an illusion of understanding arises from the possibility that since one does understand the behaviour of the parts and consequently experiences a sense of understanding this sense of understanding is taken to indicate that one also understands the behaviour of the whole. However, it is well known that unless the system-level property is exceptionally well-behaving (for example is nearly aggregative in William Wimsatt's (2000) sense), one cannot simply infer from the properties of the parts to a property of the whole.

Model builders themselves are probably relatively immune to this illusion. In fact, the frequent talk about emergence may be seen as a placeholder for the acknowledged lack of understanding of why

[5] Also, in principle, every simulation run is just a long deduction and hence understandable by a cognitive agent with sufficient working memory.

the whole behaves as it does even when the behavior of the parts is understood. The problem becomes more acute when knowledge and understanding are transferred and assessed across disciplinary boundaries.

3.2. Visualization confused with insight

Paul Humphreys, Johannes Lenhard and Eric Winsberg have all recently emphasized the importance of visualization in rendering simulation results understandable. There is no doubt that additional visual representations of the simulation results are an efficient cognitive aid in the creation of understanding (e.g. Herbert and Bell 1997): we are visually oriented creatures and it makes sense to utilize also those cognitive capacities that involve sight and spatial comprehension. However, it should be noted that merely looking at colorful pictures or animations may not by itself provide grounds for making any additional counterfactual inferences, answers to more what-if-things-had-been-different questions, about the simulated phenomenon. Conflating knowledge of dynamics with knowledge of underlying causes can be a source of illusion of understanding in the context of analytic models as well, but “seeing how the system works”, preferably in vivid color, probably enhances the sense of understanding. To see this consider a model that can be represented both as an analytic model and as a simulation: the Lotka-Volterra model.

The standard Lotka-Volterra model is a pair of first-order, non-linear differential equations that are used to describe the population dynamics of a predator and its prey. The two equations are:

$dx/dt = x(a-by)$

$dy/dt = -y(c-dx)$

In which the *x* is the hare population and *y* is the lynx population, *a* tells us how fast the hare population grows, *b* tells us how efficiently the hares are eaten up by the lynxes, *c* tells us how fast the lynxes reproduce and *d* tells how how fast the lynxes perish. In this form, the model has no analytic solution, but it is relatively easy to explore the dynamics numerically (even without a computer) and the dynamics are usually represented graphically. The graph shows how the population trajectories oscillate and the reason for this systemic behavior is easy to understand: over-predation leads to a fall in the prey-population, which leads to a fall in the predator population, which makes it possible for the prey-population to grow and so forth.

The essence of the Lotka-Volterra model can also be simulated with a cellular automata model. With a running cellular automata model, one can see how local over-predation leads to disappearance of the pray which leads to decline in the predator population and so forth. Because each state of a given cell on the screen can now be given an interpretation as a "concrete", though artificial, hare, lynx or pasture (especially if the cell states are given appropriate color codes), the simulation creates a feeling of seeing the dynamics in action, not just as an abstract representation. There is a sense of understanding of the abstract oscillatory dynamics. However, no new what-if-things-had-been-different questions (at least relating to non-spatial issues, of which the original model is silent) can be answered about the population dynamics on the basis of the cellular automata model. Hence, there is no increase in understanding proper, only an illusion of enhanced understanding.

As with the previous worry, this illusion is probably an issue mostly in contexts in which the visualizations are presented to a non-expert audience, who lack knowledge of the required context specific benchmarks and previous encounters with the kind of data that would be required for competent self-assessment of understanding. Neither is the illusion of understanding created by pictures limited to simulation results. For example, adding irrelevant brain pictures to a neuroscientific research report makes the report more convincing for non-experts and even students of the relevant field (Skolnick Weisberg et. al. 2008).

3.3. Manipulability of dynamics confused with understanding of mechanism

Johannes Lenhard (2006) argues that because of epistemic opacity of simulations, traditional (D-N style) conceptions of explanation and understanding are inapplicable. In their place, Lenhard advocates “a pragmatic conception” of understanding based on our practical ability to manipulate the simulated phenomenon according to what we have learned from manipulating the simulation model. This fits well with the conception of understanding as akin to an ability advocated in the previous section, but it is nonetheless crucial to explicate just what is and what is not understood when only the results of changes in the initial conditions or parameter values are known.

Knowledge of dependencies between inputs and outputs provides understanding of why some the end-results (the outputs) are the way they are rather than something else. For example, let us suppose that some agent based model evolves into an equilibrium E with certain characteristic C and that characteristic is dependent on the parameter value P. If we had found this out, for example by performing multiple runs with different parameter values, we would have (limited) understanding about why E has property C rather than C’. However, unless we know how that dependency itself is dependent on the structural features of the model, say that if the agents obeyed a slightly different decision rule, then the dependency between P and C would be different in some predictable way, we would not have understanding of why the system behaves as it does, i.e., understanding of the mechanism. By manipulating the initial settings we may be able to get the simulation to do what we want it to do, but this is different from understanding *why* it does so.

It is important to notice that this is a different worry from the obvious underdetermination problem that many possible causal mechanisms could produce the same observable results (in the context of simulations the underdetermination problem is often called equifinality). The illusion of understanding arising from the manipulability of the system can arise even when the simulation does capture the right causal constituents of the simulated system. If the simulation is based on totally wrongheaded causal assumptions, the understanding it provides is simply false.

3.4. The idea of maker's knowledge

Fred Dretske once asserted (1994) that in order to really understand a system, one should possess the knowledge that would, in principle, enable one to build the system. Also Joshua Epstein's motto "If you didn't grow it, you didn't explain it" (1999) could be (although perhaps slightly uncharitably) be interpreted as an expression of the same idea. As was noted above, there is a sense in which assembling a working system guarantees at least some proficiency in answering what-if-things-had-been-different questions and thus understanding. However, this understanding should be distinguished from the common but unfounded intuition of maker's knowledge: a special privileged epistemic access a maker has into his or her creation.

One of the first things to do when verifying or validating a simulation model is to see whether the program can reproduce some results known to be true of or similar to the simulated phenomenon (benchmarking). This is not an easy task. As was already noted, the sense of satisfaction in getting the computer do what you wanted it to do can be confused with understanding what is happening. Simulations also feel more concrete than analytic models. A simulation model is an artificial physical system, ecology, economy or society with which the researcher can experiment (cf. Peck 2008). Thus it seems like something that the researcher has actually built, rather than a set of stipulated assumptions and a conclusion deduced from them. However, since understanding is ultimately a matter of correct inferences concerning the effects of counterfactual changes, if the creator cannot make these inferences, no amount of spilled blood, sweat or tears can make the creator understand *why* the creation behaves as it does.

4. Improving our understanding

Since the sense of understanding is in general unreliable, and in simulation context possibly biased, indicator of understanding, we ought to replace it with something else as the central metacognitive criterion of *assessment*. As was already noted, the sense of understanding plays a crucial role in motivating research and providing immediate heuristic cues by which the model building is guided. Thus it may not be possible or even sensible to try to altogether ignore the sense of understanding during the process of model building. Since what is central to understanding is the ability to make inferences about the effect of local changes in the system under study, what can be explained by what, explicating the dependencies that ground these inferences also explicate the kind and degree of understanding. Thus the kind and degree of understanding can be explicated by stating the known relations of explanatory relevance. This should be the fundamental criterion against which models should be assessed. Explicating relationships of explanatory relevance can also be an effective strategy in *calibrating* the sense of understanding: reflecting whether one can actually answer any new w-questions concerning some puzzling aspect of a simulation or the simulated phenomenon can be used as a quick check on whether any understanding has been created.

The trouble is that when simulations become epistemically opaque, relations of explanatory relevance become hard or impossible to formulate for a cognitively unaided human. If it is indeed the case that simulation models that break the complexity barrier provide only limited understanding, how should we try to incorporate them into our overall scientific understanding of the world? First, we might try to build humanly tractable models about the simulations themselves. This strategy of *metamodelling* is an approach sometimes taken in agent based computational economics, for example. Second, we might simply have to rethink the place of the individual human mind in our collective scientific endeavor.

4.1. Metamodels

In order for a complex simulation to be understandable for a cognitively unaided human individual, it needs to be described in such a way that inferences concerning hypothetical local changes become feasible. Thus one way of improving understanding about simulations is to build additional explanatory representations, *metamodels,* about them. When the interaction of system parts becomes intractable, one way to gain some inferential power concerning possible paths of development for the system as a whole is to build a new representation of the simulation at a higher level of abstraction. As an example, Cosma Rohilla Shalizi and Christopher Moore (2003) claim that the way to create understanding of complex bottom-up simulations is to throw away micro-information in such a way, that the information left describes (in the sense of enabling to make distinctions between) macro-states with Markovian dynamics, i.e. the behaviour of the cellular automata system is described in such a level, that the future state of the system is independent of its past when conditioned on the present state. (Shalizi & Moore 2003, 10) Markov property can be naturally seen as a kind of modularity condition for stochastic processes, since it enables inferences concerning the effects of temporally localized changes in the process. In order to gain some inferential power concerning possible paths of development, we need a new representation of the CA at a higher level of description, the Markovian stochastic process. This stochastic process can be considered as an explanatory model of the original CA, which, although a model in itself, is also conceived as a system to be modelled in its own right.

The building of "empirical" metamodels[6] using standard statistical techniques on the data generated by the simulation is standard practise in some fields using micro-simulation, such as agent-based computational economics (ACE) (Kleijnen & Sargent 2000). For example, Bunn and Oliveira (2001) construct an ACE model of a wholesale electricity market to explore the possible effects of the New Electricity Trading Arrangements (NETA) introduced in the United Kingdom in March 2001. Their model incorporates strategically interacting market participants (electricity generators and energy purchasers for end-use customers); a system operator; interactions between a bilateral market, a balancing mechanism, and a settlement process; determination of day-ahead mark-ups on previous day price offers by means of reinforcement learning; and daily dynamic constraints. The result is an enormously complicated repeated stochastic game. In order to make some sense of their results, they use their simulation data to fit a number of simple econometric models describing the characteristics of market equilibria under NETA as functions of both market structure and agent characteristics. As with the cellular automata models, the mere possibility of recreating some macro phenomenon of interest from micro-foundations is not enough for understanding of the phenomenon to be created. What is required is a representation that enables unaided answers to what-if-things-had-been-different questions concerning hypothetical changes in the values of variables or parameters.

4.2. Extending the understanding subject

[6] Providing understanding of the simulated phenomena is not the only aim of metamodelling. Of independent interest may be the brute behaviour of the output, calibration or sensitivity/robustness analysis. Consequently, these different aims entail different metamodelling strategies (Kleijnen & Sargent 2000). Notice that the term metamodel is sometimes also used to refer to a kind of metatheory, according to which the simulation should be constructed and carried out.

There is no a priori assurance that all the pragmatically and theoretically important systems could be captured in such representations that would allow a cognitive agent with human limitations to reliably make counterfactual inferences about them.[7] Adding new layers of representation, models of models, may not also always be the most sensible thing to do. So far this essay has been about understanding as possessed by individual human beings. However, limiting the proper place of inferential activity to the mind of a lone heroic theorist might simply be misguided, since a great part of our cognitive practices are in any case best seen as distributed outside our minds and bodies (see e.g…). In a very important sense, we can understand the world better not because we have become smarter, but because we have cumulatively made our environment smarter. The question of the proper unit of cognition is especially pertinent in the case of scientific understanding, which is both massively distributed (within research groups and across the scientific community) and massively extended.

For Paul Humphreys (2004; forthcoming), the most important philosophical question that computational science in general and simulation in particular pose is whether we should rethink the very anthropocentric enlightenment conception of epistemology, i.e., should we let go of the presupposition that the individual human mind is the primary or default cognitive/epistemological subject. Of course, the idea of extended cognition is now generally accepted, but there is a specific point about the sense of understanding that is worth making here: the sense of understanding has probably been one major motivation for unreflectively presuming that it is the individual mind that should be taken as the seat of knowledge and understanding. For Descartes, the sense of understanding even acted as a foundational epistemic principle: the things that we understand most clearly should be taken as the most secure basis of all knowledge. Our epistemic activities are to a

[7] Herbert Simon famously argued that most evolved or designed complex systems are likely to be modular and hierarchical (Simon 1962). However, see also Kashtan and Alon 2005 for some crucial limitations to Simon’s argument.

large degree motivated by the psychological sense of satisfaction accompanying understanding, but it is still a mistake to confuse this sensation for the ultimate goal itself.

The deflationist and anti-mentalistic conception of understanding advocated here thus supports the idea that epistemological anthropocentrism should be, if not discarded, at least weakened. If the sense of understanding itself has no epistemic value, then we cannot use it to argue that attributions of understanding should be limited to conscious minds. In fact, unless there is some epistemic reason to think otherwise, the question of whether an extended cognitive system of a mind and a computer can be said to understand something even when the unaided mind is incapable of doing it becomes mostly definitional. In such a case, the only epistemically relevant facts of the matter are that the extended system can reliably answer a range of what-if-things-had-been-different questions about the simulated phenomenon, can successfully infer and explain, but the constrained cognitive system (the human) cannot. Insofar as the human-computer pair is reliably integrated to the appropriate scientific community (for example the computer and the code can be subjected to effective error control that is independent of particular simulation results), whether the extended system "really" understands or not becomes a non-issue. The human may not understand the simulation, but the human-computer pair may understand the simulated phenomenon. In fact, as was briefly noticed above, we already attribute understanding to extended cognitive systems, since even traditional analytic modeling, which essentially involves the use of pen and paper, should also be seen as extended cognition (Giere 2002, Kuorikoski and Lehtinen 2009).

The extent that understanding is allowed to seep outside our skulls varies across scientific disciplines and probably reflects important and deeply ingrained differences in methodological presuppositions and epistemic situation. Physicists do not seem to be worried about offloading cognition to machines, probably because of their relative confidence in their basic theory (from which the simulation assumptions are derived), familiarity with technology and heavy computation and, most importantly, because they have no choice. Economists are less sanguine, probably because the underlying theory is not that strong and the investigated systems are heterogeneous and constantly changing thus creating the need for general and robust models. Many theoretical economists also conceptualize their research as economic “thinking” and thus uphold the romantic image of the lone heroic theoretician, unraveling the secrets of society within his or her mind (the contentious history of the epistemic importance of introspection probably also plays a role).

REFERENCES

Baker, G. P. and Hacker, P. M. S. 2005: *Wittgenstein: Understanding and Meaning, Part I: Essays. 2nd. revised ed.,* Oxford: Blackwell.

Barnes, Eric 1992: Explanatory Unification and Scientific Understanding, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1992, I: 3-12.

Bechtel, William and Robert C. Richardson 1993: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* Princeton, NJ: Princeton University Press.

Bunn, Derek W. and Fernando S. Oliveira 2001: Agent-Based Simulation—An Application to the New Electricity Trading Arrangements of England and Wales, *IEEE Transactions on Evolutionary Computation* vol.5, no. 5: 493-503.

Cummins, Robert, Pierre Poirier and Martin Roth 2004: Epistemological Strata and the Rules of Right Reason, *Synthese* 141: 287-331.

De Regt, Henk W. and Dieks, Dennis 2005: A Contextual Approach to Scientific Understanding, *Synthese* 144:1: 137 – 170.

Dretske, Fred 1994: If You Can't Make One, You Don't Know How It Works, in P. French, T. Uehling and H. Wettstein (eds.): *Midwest Studies in Philosophy*, vol. 19: 468-482.

Elgin, Catherine 2007: Understanding and the facts, *Philosophical Studies* 132: 33-42.

Epstein, Joshua 1999: Agent-based Computational Models and Generative Social Science, *Complexity* 4 (5)): 41-60.

Fay, Brian 2003: Phenomenology and Social Inquiry: From Consciousness to Culture and Critique, in Turner, Stephen P. & Roth, Paul A. (eds.): *Blackwell Guide to the Philosophy of the Social Sciences*

Friedman, Michael 1974: Explanation and Scientific Understanding, *Journal of Philosophy* 71: 5-19.

Frigg, Roman and Julian Reiss (Forthcoming): The Philosophy of simulation: hot new issues or same old stew, *Synthese*.

Galison, Peter 1996: Computer Simulations and the Trading Zone, in Galison, Peter and David J. Stump (eds.): *The Disunity of Science: Boundaries, Contexts, and Power*, Stanford: Stanford University Press : 118-157.

Giere, Ronald 2002: Models as Parts of Distributive Cognitive Systems, in Magnani, Lorenzo and Nersessian, Nancy (eds.): *Model Based Reasoning: Science, Technology, Values*, Kluwer: 227-241.

Glennan, Stuart S. 2005: Modeling Mechanisms, *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 443-464.

Grimm, Stephen R. 2006: Is Understanding A Species Of Knowledge, *British Journal for the Philosophy of Science* 57: 515-535.

Hartmann, Stephan 1996: The World as a Process: Simulations in the Natural and Social Sciences, in: R. Hegselmann et al. (eds.), *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, Theory and Decision Library. Dordrecht: Kluwer 1996, 77-100.

Hegarty, Mary 1992: Mental Animation: Inferring Motion from Static Diagrams of Mechanical Systems, *Journal of Experimental Psychology – Learning, Memory and Cognition* 18: 1084–1102.

Hegarty, Mary 2004: Mechanical reasoning by mental simulation, *Trends in Cognitive Sciences* Vol.8 (6): 280-285.

Herbert, R.D and R.D. Bell 1997: Visualisation in the Simulation and Control of Economic Models, *Computational Economics* 10: 107-118.

Humphreys, Paul 1993: Greater Unification Equals Greater Understanding?, *Analysis* 53(3): 183–188.

Humphreys, Paul 2004: *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*, Oxford: Oxford University Press.

Humphreys, Paul (forthcoming): The philosophical novelty of computer simulation methods, *Synthese.*

Kashtan, Nadav and Uri Alon 2005**:** Spontaneous evolution of modularity and network motifs, *PNAS* vol. 102 (39): 13773-13778.

Keil, Frank C. 2003: Folkscience: coarse interpretations of a complex reality, *Trends in Cognitive Sciences* Vol.7 No.8: 368-373.

Kitcher, Philip 1989: Explanatory Unification and the Causal Structure of the World, in Kitcher and Salmon (eds.): *Scientific Explanation*, Minneapolis: University of Minnesota Press: 410-505.

Kleijnen, J.P.C. and R. G. Sargent: A methodology for the fitting and validation of metamodels in simulation, *European Journal of Operational Research* 120 (1), 14–29.

Kuorikoski, Jaakko and Aki Lehtinen 2009: Incredible Worlds, Credible Results, *Erkenntnis* 70.

Lehtinen, Aki and Jaakko Kuorikoski 2007: Computing the Perfect Model: Why Do Economists Shun Simulation?, *Philosophy of Science* 74: 304-329.

Lenhard, Johannes 2006: Surprised by a Nanowire: Simulation, Control, and Understanding, *Philosophy of Science* 2006: 605-616.

Magnani, Lorenzo and Nersessian, Nancy J. 2002: *Model Based Reasoning: Science, Thechnology, Values*, New York: Kluwer Academic/ Plenum Press.

Mills Candice M. and Keil, Frank C. 2004: Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth, *Journal of Experimental Child Psychology* 87, 1-32.

Newton-Smith, W. H. 2000: 'Explanation', in W. H. Newton-Smith (ed.), *A Companion to the Philosophy of Science*, Oxford, Blackwell: 127–133.

Peck, Steven L. 2008: The hermeneutics of ecological simulation, *Biology and Philosophy* 23: 383-402.

Rozenblit, L.R. and Keil, F.C. 2002: The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science* 26, 521–562

Salmon, Wesley 1984: *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Schurz, Gerhard and Lambert, Karel 1994: Outline of a Theory of Scientific Understanding, *Synthese* 101: 65-120.

Scriven, Michael 1962: "Explanation, Predictions, and Laws" in H. Feigl & G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, Minneapolis: 170-230.

Shalizi, Cosma Rohilla and Moore, Cristopher 2003: What Is a Macrostate? Subjective Observations and Objective Dynamics, *PhilSci Archive*, http://philsci-archive.pitt.edu/archive/00001119/

Simon, Herbert 1988 [1962]: The Architecture of Complexity, originally in *Proceedings of the American Philosophical Society*, now in *The Sciences of the Artificial, 2. ed.* Cambridge MA: The MIT Press.

Skolnick Weisberg, Deena, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson and Jeremy R. Gray 2008: The Seductive Allure of Neuroscience Explanations, *Journal of Cognitive Neuroscience* 20: 470-477.

Teller, Paul 2001: Twilight of the Perfect Model Model, *Erkenntnis* 55: 393-415.

Trout, J. D. 2002: Scientific explanation and the sense of understanding, *Philosophy of Science* 69: 212—233.

Trout, J. D. 2004: Paying the Price for a Theory of Explanation: De Regt’s Discussion on Trout, *Philosophy of Science* 72: 198-208.

Waskan, Jonathan 2008: Knowledge of Counterfactual Interventions through Cognitive Models of Mechanisms, *International Studies in the Philosophy of Science* 22: 259-275.

Wimsatt, William 2000: Emergence as Non-Aggregativity and the Biases of Reductionism, *Foundations of Science* 5: 269-297.

Wittgenstein, Ludwig 1953 [1997]: *Philosophical Investigations* (translated by G. E. M. Anscombe) Oxford : Blackwell.

Woodward, James 2003: *Making Things Happen*, Oxford and New York: Oxford University Press.

Ylikoski, Petri (forthcoming): Illusions in Scientific Understanding, In De Regt, Leonelli & Eigner (eds.) *Scientific Understanding: Philosophical Perspectives*, Pittsburgh University Press.

Ylikoski, Petri and Jaakko Kuorikoski (forthcoming): Dissecting Explanatory Power, *Philosophical Studies.*