This is the penultimate version of a paper forthcoming in Paul Humphreys & Cyrille Imbert (eds.), *Models, Simulations, and Representations*, Routledge 2011. Final version available upon request. Please cite only from the final/published version.

Scientific models, simulation, and the experimenter's regress

Axel Gelfert (National University of Singapore)

Abstract

According to the 'experimenter's regress', disputes about the validity of experimental results cannot be closed by objective facts because no conclusive criteria other than the outcome of the experiment itself exist for deciding whether the experimental apparatus was functioning properly or not. Given the frequent characterization of simulations as 'computer experiments', one might worry that an analogous regress arises for computer simulations. The present paper analyzes the most likely scenarios where one might expect such a 'simulationist's regress' to surface, and, in doing so, discusses analogies and disanalogies between simulation and experimentation. I conclude that, on a properly broadened understanding of robustness, the practice of simulating mathematical models can be seen to have sufficient internal structure to avoid any special susceptibility to regress-like situations.

1. Introduction

In this paper, I analyze the question of whether computer simulation is, in any special way, affected by what has variously been called the 'experimenter's regress' (Collins 1985) or 'data-technique circles' (Culp 1995). Such a regress, it has been argued, may obtain when the only criterion scientists have for determining whether an experimental technique (or simulation) is 'working' is the production of 'correct' (i.e., expected) data. It may seem plausible to assume that techniques of computer simulation are especially prone to such regress-like situations, given that they are further removed from nature (in ways to be specified) than traditional experimentation. In public perception, too, there appears to be a gap between the trust that is placed in the experimental success of science, as opposed to its use of computer simulation methods (e.g., in predicting global climate change).

This paper is organized as follows. Section 2 summarizes the main idea of the *experimenter's regress*, as developed by Harry Collins (1985) on the basis of a case study in experimental astrophysics. Section 3 addresses the question of whether computer simulation can properly be thought of as a form of experimentation. Section 4 identifies three clusters of scientific questions, where one might expect regress-like situations to arise in connection with computer simulation; this is followed, in Section 5, by a reconstruction of the kind of systematic considerations that may fuel worries about what I refer to as the 'simulationist's regress'. Section 6 discusses a standard response, according to which the experimenter's regress can be dissolved whenever independent measurement techniques generate sufficiently 'robust' data. However, as I argue in Section 7, in the case of simulation this standard response is not typically available, since in actual scientific practice, computer simulation studies do not always have real system as its targets; what is being 'simulated' are (sometimes quite abstract) mathematical models. Section 8 develops a richer notion of robustness that pays special attention to the invariances (and failures of invariance) of the target as well as other rigorous results that hold for the corresponding mathematical models. I conclude, in Section 9, that while data-technique circles may sometimes pose a problem for both experimentation and simulation, the practice of simulating mathematical models has sufficient internal structure, and contributes new layers of assessment, such that it is no more prone to regress-like situations than traditional experimentation.

2. The experimenter's regress

The experimenter's regress takes the form of a challenge to the standard view of experiments as providing an objective way for testing theories and hypotheses, and is based on the observation 'that experiments, especially those on the frontiers of science, are difficult and that there is no criterion, other than the outcome, that indicates whether the difficulties have been overcome' (Collins 2005: 457). Unlike other well-known challenges to the possibility of crucial experiments (e.g., the Duhem-Quine thesis), the idea here is not that we can always save a hypothesis in the face of *disconfirming* evidence. Instead, the experimenter's regress casts doubt on the possibility of saying objectively when a given empirical finding is confirming or disconfirming evidence in the first place.

For a measurement to count as good evidence, Collins argues, we must assume that it was produced by a good instrument; but a good instrument is just one that we recognize as producing good results. This is seen as introducing a circularity: 'we won't know if we have built a good detector until we have tried it and obtained the correct outcome. But we don't know what the correct outcome is until … and so on *ad infinitum*.' (Collins 1985: 84) Controversies about the validity of empirical measurements cannot be closed by objective facts because there are no conclusive criteria 'other than the outcome of the experiment itself' that scientists could apply to decide whether the experimental apparatus was working properly or not (Collins 2005: 457). In response to two of his critics, Collins identifies as the target of the experimenter's regress the popular belief that science 'had eliminated disputes about standards in matters to do with observation of the natural world' (Collins 2002: 154). The experimenter's regress, in this sense, has an explicative role with respect to science: it is meant to show both 'what it is that is taken to be the normal standard for experimental replication – getting the right result' and the source of that belief: namely, '*the ready availability of a standard just in those cases where experimental replicability is not needed for proof*' (Collins 2002: 155; italics original).

In support of his claims, Collins draws on a detailed case study of the scientific debate surrounding the search for gravitational waves, which are predicted by general relativity theory and which would manifest themselves as minute disturbances of the measured gravitational constant. Collins analyzes how one scientist, Joseph Weber, built a detector – basically a massive piece of metal hooked up to extremely sensitive sensors that would detect any disturbance passing through the device – and then announced to the physics community that he had detected gravitational waves. Weber's claims were immediately contested by other scientists, some of whom built their own detectors (and could not corroborate the findings), whereas others disputed Weber's results by criticizing the validity of the procedures that Weber had used to calibrate his own instruments. To simplify somewhat, the dispute was never settled on objective grounds, but petered out slowly as Weber's track record as an experimenter came under fire and his reputation gradually eroded. In the end, Collins argues, when agreed-upon methods fail 'scientists turn to a variety of criteria broadly similar to those used in making common-sense judgments in ordinary life' – concerning reputation, institutional affiliation, track record, personal ties, conduct, and so forth. (Collins 2005: 458) Whether Collins's argument does indeed extend to 'the practice of disputed science' (Collins 2002: 155) in general, has been a bone of contention for some time (see, Franklin 1994, for a wholesale critique); in the present paper, I shall limit myself to the question of whether *computer simulation studies* are especially prone to versions of Collins's regress argument.

3. Simulation as 'experimenting on theories'

Simulations as routine tools of scientific inquiry are relatively more recent than scientific models; however, like the use of scientific models, the practice of computer simulation has often been located (though not unproblematically) at an intermediate level between theory and experiment or experimental data. This is evident from the way practitioners of computer simulation themselves have, from early on, described what it is they are doing. In his study of the historical origins of 'Monte Carlo' simulations in high-energy physics, Peter Galison quotes the author of a 1972 review article on computer simulation, the computational physicist Keith Roberts, as follows:

> Computational physics combines some of the features of both theory and experiment. […] It is symbolic in the sense that a program, like an algebraic formula, can handle any number of actual calculations, but each individual calculation is more nearly analogous to a single experiment or observation and provides only numerical or graphical results. (quoted in Galison 1996: 137)

In the early days of computational science, 'simulation' could be hands-on in a quite direct way. Herbert A. Simon, in his 'Comments on the History of "Simulation"', recalls how, in an early (1956) paper, he and colleague spoke of 'hand simulation' to indicate 'that the program was not yet running on the computer, but that we had hand simulated its processes and thought we could reliably predict its behavior'. The irony of then manually implementing what would nowadays be an automated procedure is not lost on Simon: 'Here we [were] talking of people simulating a computer rather than a computer simulating people!'[1]

At a descriptive level, computer simulationists share many of the procedures of experimentalists, among them 'a shared concern with error tracking, locality,

[1] Quoted in (Röller 2008: 52-53).

replicability, and stability' (Galison 1996: 142), as well the initial challenge of 'getting the simulation to behave properly at all' (Kennefick 2000: 26). It is not at all unusual to come across, especially in the writings of early computational physicists, references to computer simulations as 'numerical experiments', 'computer experiments' and so forth.[2] While such phrases convey a sense of how the practice of computer simulations is experienced by those who engage in it, they are also philosophically vague. As Hans Radder notes, the phrase 'computer experiments' is really an umbrella term for 'various sorts of hybrids of material intervention, computer simulation, and theoretical and mathematical modeling techniques' (Radder 2005: 271). As such, it stands in need of elaboration, and the conceptual relations between computer simulation and (traditional forms of) experimentation need to be investigated.

One bone of contention is the significance of materiality in traditional experimentation as opposed to the (allegedly) more ephemeral character of computer simulations. It has been variously suggested that computer simulation fails to meet certain criteria associated with 'proper' experimentation. Thus Francesco Guala (2005: 214) argues that, whereas in traditional experimentation the same material causes operate ('at a "deep", "material" level') in both the experimental and target systems, in computer simulations one can at best expect an 'only abstract and formal' correspondence relation (of similarity or analogy) between the simulating and the target systems. However, as Wendy Parker points out, while *material* similarity may often allow us to extract information about the target system, it is only one dimension along which one can assess the power of experiments and simulations to enable inferences about the target system. As a case in point, she refers to the example of weather forecasting. Re-creating a system that is materially continuous with the target system – that is made 'of the same stuff', i.e. clouds, water, mountains – is rarely feasible, nor is it very promising for the purpose of day-to-day weather forecasts. (Parker 2009: 492) By running computer simulations of the underlying weather dynamics, scientists are in a much better position to make warranted predictions about tomorrow's weather. Hence, it need not be generally the case that, as some authors have suggested, traditional experiments 'have greater potential to make strong inferences back to the world' (Morrison 2005: 317).

[2] See, for example, the chapter on 'The Computer Experiment' in (Hockney and Eastwood 1988), and references therein.

The tendency to think of computer simulations as, in some fundamental sense, more ‘abstract’ than traditional experimentation, may be partly due to an ambiguity in the term ‘computer simulation’ itself. If one understands by ‘computer simulation’ the successful implementation of a computational template, which is then executed to generate a set of simulated data, then each run may be considered a new simulation. Assuming that the actual device, on which the simulation is being run, works correctly, there does not seem to be much room for questions of validity (and the corresponding ‘shared concerns’ – of error tracking, replicability, etc. – that dominate traditional experimental practice): the correct outcome is simply whatever result is generated, and typically displayed, by the computer. This picture, of course, is too narrow, since by ‘computer simulation’ one does not simply mean the outcome of any particular run of an algorithm; computer simulations are deployed in concrete settings, according to the goals of the specific project at hand, in order to investigate target systems across a range of parameter values. When thus viewed in terms of their pragmatic role in scientific inquiry, simulations, too – including the devices they run on, and the programming code – are subject to evaluation, revision, and improvement. As Parker puts it, although ‘computer simulations *per se*’ – understood in an abstract sense, as a time-ordered sequence of states – ‘do not qualify as experiments, computer simulation studies do’, where ‘computer simulation studies’ refers to the investigative activity on the part of a researcher, who uses computer-generated results as input for further inquiry. (Parker 2009: 495)

As this brief discussion suggests, characterizing computer simulations as ‘numerical experiments’ requires greater care than is typically exhibited by those practitioners of simulation studies who employ the term. Paul Humphreys may well be right when he writes that ‘claims that [computer simulation] methods lie “in between” theorizing and experimentation are […] best interpreted metaphorically’ (Humphreys 2009: 625). However, this should be understood as a *caveat*, not as a prohibition against pursuing the parallels between scientific experimentation and computer simulation, and treating simulation studies as a form of experimental practice – though a novel one that raises its own, recognizably philosophical problems.

4. Formulating the 'simulationist's regress'

If one accepts that computer simulation studies are in many ways on a par with experiments, and indeed may be thought of as continuous with experimental practice, then one should expect to encounter arguments concerning their replicability which are similar to the ones that gave rise to the formulation of the experimenter's regress. Note that, at this stage, it is not necessary to endorse any sweeping claims that the experimenter's regress does *in fact* undermine the possibility of resolving experimental disputes through further empirical inquiry and rational debate. All that is required at this point is the acknowledgment that, to the extent that replicability raises serious questions for the practice of scientific experimentation, analogous questions can be raised for 'numerical experiments', i.e. the practice of computer simulation studies.

In direct analogy with Collins's discussion, one can then define what one might call the 'simulationist's regress', as applicable to those situations *where the best or only test of a simulation is its own disputed result*. As in the case of experiments, the regress will typically become apparent only in stubborn cases of disagreement, when mutually agreed methods of resolving disputes have failed. Before turning to specific examples of how the simulationist's regress might unfold, it is helpful to adopt a bird's-eye perspective and ask for general characteristics of situations where one might expect the regress to become salient. Collins's example of the search for gravitational waves is instructive in this respect, since it combines three important characteristics, each of which may contribute to the emergence of a regress. First, the example concerns the search for an *as yet unobserved*, and in many ways *causally isolated*, phenomenon. (Gravitational waves can neither be created, nor manipulated, nor do they have any observable effects outside the context of specific purpose-built detection devices.) Second, any theoretically predicted effect would be extremely small, thereby rendering even the *existence claim* that comes with a purported observation highly controversial. (Much of the controversy about gravitational waves concerned the question of how they would manifest themselves in any proposed experiment – i.e., what their experimental 'fingerprint' would be.) Third, even on the most optimistic astronomical theories, measurable gravitational waves will be fairly rare, thus leading to a *sparsity of empirical data*. Taken together, these three factors give rise to the regress problem described by Collins: how are experimental disputes to be resolved, if the claims in question concern the existence of causally isolated (or otherwise inaccessible) processes, for which empirical evidence is sparse?

The example of gravitational waves is special, insofar as it *combines* the three aspects distinguished above, thereby making the alleged regress especially salient. In the remainder of this section, I wish to argue, first, that less salient versions of the same problem can arise even when not all three aspects are instantiated, and second, that analogous considerations apply in the case of computer simulations. For one thing, simulations often take place under conditions of sparsity of empirical data with which to compare the simulation results – either because such data is not readily available, or because it cannot be generated at will in a controlled experiment. As Eric Winsberg points out: '[S]imulations are often performed to learn about systems for which data are sparse. As such, comparison with real data can never be the autonomous criterion by which simulation results can be judged.' (Winsberg 1999: 289) Whether or not a simulation result is to be believed, is thus not generally something that can be settled by comparison with empirical data alone, at least not in cases of genuine dispute about the validity of the simulation itself.

In order to illustrate how regress-like worries can arise in the case of computer simulations, I want to describe briefly, without getting sidetracked by the nitty-gritty detail of a complete case study, three classes of actual scenarios, in which the best test of a simulation may be its own disputed result. Each example corresponds to one of the three aspects distinguished earlier, that is *causal inaccessibility*, *novelty of the phenomenon* (which issues in an *existence claim*), and *sparsity of* (real, empirical) *data*.

The first scenario is based on an existing case study from astrophysics (again!), and concerns simulation-based studies of the collapse of neutron stars. It is perhaps no coincidence that astrophysics is a rich source of examples, given that the target systems under investigation are often causally remote, in the sense that they do not lend themselves to manipulation or causal intervention in a controlled experiment. In his detailed case study, Daniel Kennefick (2000) reconstructs the controversy between different groups of scientists working on the astronomical analysis – and computation simulation – of how neutron stars may, given certain conditions, form binaries (i.e., systems of two neutron starts orbiting one another), which eventually collapse into one another, thereby forming a black hole. Specifically, Kennefick analyses the reaction of the astronomical community to the claims made by a pair of computational physicists, who thought they had identified, through the use of computer simulation, a scenario in which one of the neutron stars forms a black hole – through a process Kennefick dubs 'star crushing' – *before* it swallows up the other one. For various reasons, this claim contradicted what most analytically-minded theorists of general relativity expected, who immediately criticized the simulation results (in sometimes a rather wholesale fashion) as erroneous and due to programming errors. Focusing his attention on the divergent methodologies (and their corresponding criteria of what constitutes a valid derivation) of relativity theorists and computational physicists, Kennefick argues that theoretical predictions are subject to a 'theoretician's regress', which arises from the 'near impossibility of discriminating between the results of rival calculations which fail to agree, *by the process of criticizing the calculations*' (Kennefick 2000: 33-34; italics added).

The second scenario in which one might expect regress-like situations to emerge, can be illustrated by the simulation-aided search for the presence, or absence, of phase transitions in complex systems. While phase transitions are among the most salient phenomena in nature – one need only think of the freezing of water, the occurrence of spontaneous magnetization in certain metals, or the phenomenon of superconductivity – physicists are hard-pressed to come up with theoretical models that successfully explain their occurrence. The study of many-body models poses complex numerical challenges, and the full set of equations describing a many-body system is almost never analytically solvable. This has led to a proliferation of simplified mathematical models – such as the Ising model or the Heisenberg model – to capture the 'essential physics' that is thought to be responsible for the phase transition.[3] In order to evaluate such models, and whether they can account for phase transitions, computer simulations are being run to determine whether there is a region in parameter space where a phase transition occurs. The occurrence of a phase transition is typically indicated by the asymptotic behavior of certain variables (such as the correlation length, or a suitably defined order parameter) as the system approaches the critical point. For most many-body models in three dimensions it is unknown whether they are indeed capable of reproducing phase transitions; as a result, the search for models that have this capacity takes the form of a search for regions in parameter space where the relevant variables display 'critical' behavior.

[3] On this point see also (Batterman 2002).

Hence, in the case of a purported positive result, we are faced with an *existence claim* – 'that there *is* a phase transition in *this* region of parameter space, which matches the kind of phase transition in the target system' – where the best evidence for the success of a simulation is its own disputed result.

For the third kind of situation where one might expect to find a version of the simulationist's regress, I want to point to large-scale simulations arising within what is sometimes called Earth System Analysis, such as global climate models.[4] At first it might seem counterintuitive to lament a '*sparsity* of empirical data', given that one challenge of simulating global climate change consists in integrating a diverse range of variables and complex data sets, and the relations between them. What is important, however, is not the actual *amount* of data involved, but the availability of *independent* data sets for the same variables. For example, one would typically like to compare the outcome of different simulation runs for different parameter values (e.g., certain levels of greenhouse gas concentrations), against experimental data corresponding to these different situations. However, when it comes to climate models, no independent experimental access to the target system – the Earth's climate – is possible: the only experiment that takes place is the irreversible climate change that we, as humans, are currently inflicting on the planet and its geochemical cycles. It may sometimes be possible to turn to substitutes for independent experimental 'runs': in the case of climate modeling, for example, one might simulate different episodes in the Earth's climate and compare the simulation results with historical climate records. While it may sometimes be an option to run simulations of past, observed climate anomalies in order to test one's numerical models, one should not necessarily expect this to resolve any disputes unless there is a prior agreement about the – typically *sparse* – data that goes into any given simulation. In cases where such agreement is lacking, one can easily imagine scenarios where the validity of the corresponding results is challenged – on the basis of worries about the admissibility of past evidence, or by contesting the independence of the relevant data subsets – thereby rendering disputes about the significance of simulation results irresolvable.

[4] For a different historical example, see Peter Imhof's study of the uses of computer simulation in the Limits of Growth controversy (Imhof 2000).

5. Anatomy of a regress

Having identified scenarios from different areas of research where one might expect regress-like situations to occur in computer simulation studies, let us approach the problem from a more systematic point of view, before discussing possible responses to it. Rather than ask which scientific questions or disciplines might be particularly susceptible to the simulationist's regress, I want to refocus attention on its similarities and dissimilarities with the experimenter's regress. I shall do so by developing two complementary perspectives on practical side of computer simulations, relating in turn to the software and hardware aspects of their implementation.

Numerical techniques and computing power have become important drivers of research across many scientific disciplines. Heavily-tested numerical software packages have turned computational methods into a commodity that is available to a wide range of researchers. Here is how two researchers in metabolic drug design describe the situation in their discipline: 'Modern programs such as *Gepasi* (Mendes, 1993) and *SCAMP* (Sauro, 1993) are easy to use, run on universally available equipment, and powerful enough to handle most of the problems likely to interest the metabolic simulator.' (Cornish-Bowden and Eisenthal 2000: 165-166) While over-reliance on pre-packaged computational software may lead to epistemic problems of its own, regarding how to assess the validity of the results its generated, it is important to note a crucial contrast with possible cases of the simulationist's regress – such as Kennefick's example of 'star crushing'. In cutting-edge science – especially where it operates under conditions of sparsity of data, or deals with existence claims and situations of causal isolation – simulations are typically purpose-built and tailored to specific theoretical scenarios and situations. Thus, in the 'star crushing' example, Wilson's and Mathews's results were attacked by a majority of relativity theorists and computational astrophysicists, not because of doubts concerning the validity of those approximations that were regarded as standard – as Kennefick notes, a number of elements in Wilsons's and Mathews's simulation 'resemble[d] tricks commonly used for many years in analytic GR [=general relativity]' (Kennefick 2000: 14) – but because of theoretical considerations that made the alleged crushing effect unlikely, along with a subsequent failure of other numerical groups to replicate the effect. When replication fails (or is not even attempted because the result is deemed erroneous for other reasons), the proponent of the disputed simulation result faces a similar challenge to the experimenter, in that ‘the experience of making a calculation work, which is part of what convinces the theorist that the result can be believed, is so particular to the specific theoretical apparatus and its operation that it is of little help in convincing those who disbelieve the result on other grounds’ (Kennefick 2000: 33).

The fundamental problem, thus, in one of the *replicability* of simulations, i.e., their ability to generate stable data. Achieving replicability has both a ‘software’ and a ‘hardware’ aspect, which relate to different levels at which failures of replicability can (and sometimes do) occur. I shall begin with a discussion of the ‘software’ aspect, which takes its lead from the earlier observation that cutting-edge simulations – of the sort that are most likely to give rise to potentially regress-like disputes – are typically tailored to specific theoretical problems or scenarios (such as the theoretical description of collapsing neutron star binaries in terms of approximations to Einstein’s field equations). In such cases, there is no agreed-upon procedure for deriving (stable and computationally tractable) computational templates from fundamental theory.[5] As Kennefick rightly points out: ‘In numerical work a great part of the effort is in getting the simulation to behave properly at all.’ (Kennefick 2000: 26) As with experiments, this stage of setting up simulations often involves tacit know-how on the part of the investigator and is rarely fully documented. Adopting Hans Radder’s terminology regarding the epistemology of experiment, one might say that what is being achieved during this phase of ‘getting the simulation to behave properly’, in various internal or external settings (i.e. for different parameter values as well as across different computers), is the mere replicability of the *material realization* of a simulation (i.e., its implementation on a concrete computing device). At this level, *replicability* – while being far from theory-free – is not yet assessed under a *specific* theoretical interpretation of the simulation results. (See Radder 1996: 17 & passim.)

The epistemic significance of this phase of ‘getting a simulation to behave properly at all’ lies not in furnishing *simulation results*, which may then be compared against empirical data or theoretical predictions. Rather, what goes on during this phase may, on occasion, preclude which results a simulation is capable of generating later on – once it has successfully been ‘tamed’, as it were. For, this phase will be considered to have been concluded only once the simulation begins to give the kinds

[5] On the notion of ‘computational template’, see (Humphreys 2004: 60-76).

of results we expect it to give. But, of course, under any realistic conditions of inquiry, we would not be needing the simulation if we already knew what to expect; hence, the best or only test of whether a simulation is working 'properly' may, once again, lie in its own disputed result.

Replicability in the case of computer simulation, however, also has a 'hardware' component, which is in addition to the difficulty, discussed so far in this section, of assessing whether a particular computational template is *in principle* capable of successfully replicating the behavior of a given target model. The additional worry concerns the further problem of the reliability of the concrete device, on which a simulation is being run. After all, any computer is itself a material piece of hardware and assessing whether it is a 'properly functioning' device – irrespective of the stability of any computational template that might happen to be implemented on it – is subject to the experimenter's regress, or so the worry goes. Whereas the 'software' aspect of computer simulation concerns the stability, tractability, and replicability of a piece of specially programmed code on a computational device that is assumed to be functioning properly, the 'hardware' aspect concerns the replicability of the material realization of a simulation (i.e., its implementation on a concrete computing device), for various internal or external settings (i.e. for different parameter values as well as across different computers). Drawing a parallel to recent work on the philosophy of experimentation (e.g., Radder 1996), one might say that, at the 'hardware' level, replicability – while being far from theory-free – is not yet assessed under a specific theoretical interpretation of the simulation results. What is at stake is not so much the adequacy or suitability of the computational template and its programmed implementation, but the proper functioning of the computing device itself. More often than not, it is taken for granted that the physical apparatus on which a simulation is run is indeed a correctly functioning computer; however, from a foundational viewpoint that aims to identify potential sources of a regress, 'the theory of the apparatus in simulation is not negligible' (Weissert 1997: 112). Even a carefully tested and designed digital computer is, by necessity, only capable of finite-precision arithmetic. This has several important implications for the replicability of simulations. For one, due to round-off errors, iterations of arithmetic operations will lead to a cumulative increase in the uncertainty of any numerical outcome. The precision of which a computing device is capable – which is in large part determined by the hardware – is especially important when dealing with regions of deterministic chaos, where minute differences can lead to very different system trajectories. This may eventually result in a loss of the true solution in the noise that is generated by the numerical procedure. As Weissert emphasizes: ‘This effect is not a loss of determinism in our dynamical system, but a loss of our ability to make the determination with a simulation.’ (Weissert 1997: 115) However, it is important not to overstate the *special* difficulties posed by finite-precision effects inherent in the hardware. For, it is possible to numerically estimate the propagation of finite-precision and round-off errors and their effect on the outcome of a simulation. This itself indicates that the kinds of challenges posed by the ‘hardware’ and ‘software’ aspects of replicability cannot always be neatly kept apart. Achieving replicability is a matter of eliminating sources of error and noise, and these include, amongst others, ‘computation instability, truncation error, iterative convergence error, programming mistakes, hardware malfunction, error in the mathematical form of the continuous model equations, error in the parameter values included in those equations etc.’ (Parker 2008: 176-177). While this list may look dispiritingly long, it also contains a grain of hope, for it suggests that, rather than being faced with a ‘nesting’ of *independent* sources of possible regress-like situations – e.g., hardware limitations, or particular features of the algorithms used in the software – it may be possible to help oneself to a range of techniques of achieving and testing replicability. Precisely because replicability is a composite phenomenon, it may be possible to assess the extent to which one ‘ingredient’, relative to all others, poses a threat to replicability; one can then hope to identify conditions under which the simulation is sufficiently robust for the investigative purposes in question. This suggests a way of defusing possible worries about the dangers of a *simulationist’s regress* – ‘on top of the experimenter’s regress’, as it were – namely by showing that, perhaps, there is nothing special (or intrinsically less tractable) about regress-like situations in the case of simulation, as compared to experimentation. What one is dealing with, in either case, is the challenge of *achieving replicability*, and while this may take different forms in different experimental or simulation contexts, it is not unreasonable to expect that new contexts may also give rise to new benchmarks or mechanisms of how the stability of one’s results can be ensured.

6. The response from robustness

Among the possible responses to 'data-technique circles', of the sort represented by the experimenter's and the simulationist's regress, two general approaches can be distinguished. First, one can begin with scientific practice, by identifying strategies that scientists employ in order to build confidence in their (experimental or simulation) results. Allan Franklin, on the basis of case studies of experimentation, has drawn up a list of such strategies (Franklin 1986; 2009), the most important of which are *comparison* (with known features of the system under investigation), *calibration* (of a given method against its performance in other systems), *verification* (of one experiment by other experiments or experimenters), *retrodiction*, *consistency*, *intervention* (including variation of input parameters), and *elimination* of likely sources of error. Several of these strategies have recently attracted attention in relation to computer simulation (Weissert 1997: 122-125; Parker 2008). I shall here, however, focus on a second, more abstract approach to the question of when we are justified in placing trust in the (non-obvious, or even controversial) results of scientific experimentation and simulation.[6] This response takes its cue from the familiar problem of when to trust what others tell us. Imagine that a hearer encounters a series of unexpected instances of testimony, each reporting the same unusual event. Should she believe what she is told, even when she has *prima facie* reason to think it implausible? It depends on whether the testifiers are independent or not. If the hearer has good reason to believe that the testifiers she encounters *really are* independent, and their testimonies concur in all relevant respects, then this should boost her belief in the truth of the testimony, even when the asserted fact is individually implausible. Moreover, it would be rational for her to believe the reports, given their unanimous concurrence, even when she knows each testifier to be *individually unreliable*. With independent pieces of evidence, the likelihood of error is the likelihood that *all* the evidence is wrong at the same time (and that all its ingredients are wrong *in exactly the same way*, given their concurrence) – and this becomes an ever more remote possibility the more concurring instances of evidence are involved. It is for this reason that police routinely accept the concurring, uncontradicted testimony of, say,

[6] It is worth emphasizing that this general approach is complementary to, and compatible with, Franklin's identification of specific strategies (in particular, Franklin's admission of *statistical* arguments into the array of confidence-building strategies).

accomplices even when they know them to be individually unreliable on most occasions – just not on this one.

Applying similar considerations to the case of 'data-technique circles', one arrives at what one might call the *response from robustness*. This response to Collins's experimenter's regress has been aptly argued by Sylvia Culp (1995), who describes the basic idea as follows:

> When comparable data can be produced by a number of techniques and the raw data interpretations for these techniques do not draw on the same theoretical presuppositions, this remarkable agreement in the data (interpreted raw data) would seem to be an improbable coincidence unless the raw data interpretations have been constrained by something other than shared theoretical presuppositions. (Culp 1995: 448)

Culp's paper takes the form of a case study of different DNA sequencing techniques, each of which exploits different parts of the overall framework of the theory of DNA. For example, one technique might exploit the effect of certain enzymes that cut DNA at particular points in the sequence, whereas the other might be based on the individual replacement of particular nucleotides with different bases. By employing different techniques, 'each of which is theory-dependent in a different way', it is possible – Culp argues – to eliminate the overall 'dependence on at least some and possibly all shared theoretical presuppositions', thus giving rise to bodies of data that have been produced by *de facto* independent techniques. (Culp 1995: 441) If these data sets all support the same theoretical interpretations, then we can accept these results with good reason, since the concurrence of independent techniques would otherwise be a near-miracle. Culp argues not only that is it possible to eliminate theory-dependence (and with it the danger of circularity) in this fashion, but additionally that 'it is the production of robust bodies of data that convinces scientists of the objectivity of raw data interpretations' (ibid.).

Following on from the empirical claim that robustness is what *actually* generates agreement among scientists, three problems with Culp's account need to be mentioned. First, there is a tension with the historical observation – impressively substantiated by Hasok Chang's recent (2004) study of the evolution of techniques of thermometry – that, in many areas of science, calibration across theoretically interdependent techniques is the norm, whereas true independence is the exception. This historical point is but the flipside of a second, more systematic consideration. What would be the alternative to calibrating a given measurement technique against existing standards and techniques that are theoretically 'of a piece', as it were, with the technique under investigation? We would be thrown back on our prior (theoretical and pre-theoretical) expectations of what a good result would be. However, these expectations are no more independent from one another than, for example, the various measurement techniques for temperature discussed in (Chang 2004). Hence, even if all our seemingly 'independent' measurement techniques were indeed calibrated only according to our (pre-)theoretical expectations, given that these expectations do not form an independent set, the alleged independence of our techniques would once again be undermined. Finally, if we *did* have independent measurement techniques that were neither calibrated against one another, nor calibrated according to the same expectations and evidence, how could we be so sure they would all be measuring the same quantity? Or, alternatively, how could disputes about whether or not a newly proposed technique does *in fact* measure the same quantity, possibly be resolved?[7] It would appear that total independence could only be achieved at the cost of incommensurability of one sort or another.

It should be noted that thinking of experimentation merely in terms of the deployment of *measurement techniques* is guaranteed to paint an impoverished picture of the nature of experiment. It is a striking feature of many of the examples that are cited in support of the experimenter's regress – most prominently, the detection of gravitational waves – that there is not actually all that much *experimenting* going on. The experiments in question tend to be cases of passive detection; they do not involve the systematic exploitation of causal powers and their 'redeployment' in other parts of nature – in short, the kind of 'intervening in nature', that, in the eyes of Ian Hacking and others, is crucial to the trust we place in experiments.[8] (Hacking 1983: 38) A similar sentiment appears to motivate Parker's emphasis that, if one is to think of computer simulation as a kind of *experimental activity*, one must go beyond its definition as a 'time-ordered sequence of states' and introduce the broader notion of *computer simulation studies*, which is taken to also include actions on the part of the simulationist, through which he intervenes in a system. (Parker 2009: 495) As I shall argue in the remainder of this paper, the notion

[7] Chang makes a similar point with respect to the historically contested technique of Wedgwood pyrometry (Chang 2004: 127).

[8] Indeed, as Franklin acknowledges, this is why any complete 'epistemology of experiment' must include specific confidence-building strategies '*along with* Hacking's [criterion of] intervention and independent confirmation' (Franklin 2009; italics added).

of robustness, too, must be broadened beyond the mere insistence on independent techniques of data-generation if it is to be of use against the simulationist's regress.

7. Models as targets

Just as one finds different views on whether or not computer simulation should properly be regarded as a kind of experimentation, one also finds divergent views on the question of what the 'target' of computer simulation really is. What does simulation typically aim at? Eric Winsberg has argued that, whereas experiments 'often play the role of providing crucial tests for theories, hypotheses, or models', simulations ordinarily cannot, since they assume significantly more background knowledge – including knowledge 'about how to build good models of the very features of the target system that we are interested in learning about' (Winsberg 2009: 587). What matters for simulations, on this account, is their external validity: how well a simulation performs with respect to the (physical) target system and its relation to the outside world. Indeed, it is the 'possession of principles deemed reliable for building models of the target systems' that, according to Winsberg, plays a criterial role in *defining simulation* in the first place: such possession of background principles is what justifies our treatment of a simulation, implemented on a digital computer, as an adequate stand-in for the target. (Winsberg 2009: 588)

It is helpful, in this context, to briefly return to the earlier distinction (see Section 3) between 'computer simulation' *simpliciter* and 'computer simulation studies'. Both often get subsumed under the umbrella term 'simulation', but differ in relevant respects. When one speaks of a simulation (*simpliciter*) of a physical system *P*, there is a clear sense in which the term 'simulation' is employed as a success term: for something to succeed – however imperfectly – in simulating *P*, it is necessary that *P* exist and that one's simulation stand in the right sort of relation to it. Typically, this is achieved through building theoretical models of the target system *P* and implementing appropriate computational templates on a computing device. Clearly, then, someone who intends to simulate a physical system *P* must possess (or at least take himself as possessing) sufficiently reliable principles for modelling *P*. This, I take it, is what lies at the heart of the claim that what distinguishes experimentation from simulation is that the latter requires 'possession of principles deemed reliable for building models of the target systems'. However, if one means by 'simulation' a general kind of *theoretical activity* (as suggested by the expression 'computer simulation studies'), rather than the establishment of a *relation* between a computational procedure and a particular physical system *P*, then more exploratory uses of simulation might be legitimate, even in the absence of specific theoretical knowledge about any particular target system. In the remainder of this section, I want to suggest that *computer simulation studies* need not, and indeed cannot, always presuppose a good prior grasp of what makes something a good model. To be sure, in many cases it may be possible to assess the success of simulations directly in light of their external validity against experiments and observations, especially when there is agreement about which model properly represents the target systems (think of a Newtonian model of the solar system). However, when there is a lack of such agreement, it quickly becomes apparent that, as Margaret Morrison puts it, 'strictly speaking the model is what is being investigated and manipulated in a simulation' (Morrison 2009: 45), not the target system. It is in this sense that one can – as scientists often do – speak of 'simulating a model'. Simulations do not always have real physical (or biological, or otherwise materially constituted) systems as their immediate target. Instead, in many cases, what is being simulated are mathematical models, i.e. abstract objects that are not themselves (and do not purport to be) part of the causal fabric that a physical experiment would typically exploit.[9] That many of the most widely investigated – and in this sense 'successful' – models in science are, strictly speaking, uninstantiated is hardly news to anyone who is aware of the role that idealization and abstraction play in their construction. But models are often investigated not with an eye to how accurately they represent any particular target system, but with an eye to whether or not they exhibit certain theoretically expected behaviors – as in the case of phase transitions mentioned above. In many scientific contexts, one is less interested in a particular model with fixed parameter values than in a class of models defined by some (continuously varying) control parameter. In such cases, it is often 'the entire model class that becomes the entity class that becomes the entity under scrutiny' (Weissert 1997: 108). The *real* systems, in which the phenomenon of interest has been observed (e.g. spontaneous symmetry-breaking, in the case of phase transitions), need not individually be accurately described by the

[9] Note that this is not meant as a *criterion* for distinguishing between simulation and experiment.

mathematical models in question. In fact, the class of real systems can be so diverse that it would be quite misleading to jump to conclusions about the external validity of simulation results for any one subclass of real target systems.

Simulation, thus, is in many ways a less constrained practice of scientific investigation than causal-interventionist experimentation, making it more difficult to tell whether an unexpected result represents a feature of a 'real' target system or is a mere artifact. Galison makes a similar observation when he writes:

> From the start, simulations presented a hybrid problem. On one side, the work was unattached to physical objects and appeared to be as transportable as Einstein's derivation of the A and B coefficients for quantum emission and absorption. But in practice, this was hardly the case. (Galison 1996: 140)

Interestingly, computer simulationists have sometimes taken this causal 'detachedness' of simulations to be an advantage: Galison quotes Herman Kahn, of the RAND corporation, as saying that Monte Carlo simulation methods 'are more useful […] than experiments, since there exists the certainty that the comparison of Monte Carlo and analytic results are based on the same physical data and assumptions.' (Galison 1996: 143) However, unless more is said about why dependence on the same theoretical assumptions is unproblematic, rather than a compounding factor in any looming data-technique circles, such optimism is little more than a leap of faith. In the remainder of this paper, I wish to develop a position that acknowledges (and endorses) a certain degree of independence of simulation from experiment, while also providing the internal resources to fight off the danger of vicious data-technique circles that might arise from the fact that 'the assumptions of the underlying model are, barring incompetence, built into the simulation and so the simulation is guaranteed to reproduce the results that the model says it should' (Humphreys 2004: 134). In order to do so, it is important not to assimilate questions about the trustworthiness of simulation to questions about the external validity of particular simulation results, but to recognize that simulation – in particular, in relation to mathematical models – introduces new dimensions along which its trustworthiness can be assessed.

8. Robustness, invariance, and rigorous results

Computer simulation stands in need of justification, and the same confidence-building strategies that have been discussed for experimentation – calibration, verification, elimination of error, comparison with known results etc. (cf. Franklin 1986, Parker 2008) – can also be extended to the case of simulation. In particular, as Winsberg puts it, 'the first criterion that a simulation must meet is to be able to reproduce known analytical results' (Winsberg 1999: 189). Likewise, Humphreys writes: 'In the case of computational devices, the calibration standards that must be reproduced are analytically derived results which serve as mathematical reference points.' (Humphreys 2004: 117) The significance of such analytical results as 'mathematical reference points', I want to argue, plays a special role in computer simulation, inasmuch as it goes beyond the mere reproduction of certain numerical values that are known to obtain (e.g., for certain parameter values), and beyond their uses in calibrating the computer *qua* physical device. Analytical results, in the form of what first came to be known in statistical physics as *rigorous results* (see Ruelle 1969 and Gelfert 2005), add structure to mathematical models that can neither be deduced from fundamental theory nor inferred from (or, for that matter, compared with) experimental data. In addition to exact solutions – which, for many mathematical models, are hard to come by – other examples of rigorous results might include impossibility theorems, upper and lower bounds for certain variables, or mappings between different mathematical models. The latter – the existence of rigorous results that relate different classes of models – has also been shown to lead to qualitatively new ways of indirect confirmation of models (and *across* models). Thus, in one recent example from condensed matter physics, it has been argued that it is only in virtue of a mathematically rigorous mapping between two mathematical many-body models -- the Hubbard model (for itinerant electrons) and the Heisenberg model (for fixed spins) -- that the former was recognized as successfully representing a particular class of real, experimentally known physical systems (known as Mott insulators). (See Gelfert 2009: 509-517, for a detailed study of this example.)

It should come as no surprise, then, that such rigorous results and relations have been used by scientists for the calibration of computer simulation methods: 'The partial independence of rigorous results from fundamental theory, and the fact that they are model-specific, makes them interesting "benchmarks" for the numerical and analytical techniques of calculating observable quantities from the model.' (Gelfert 2009: 506) However, the mere existence of a benchmark does not guarantee that a simulation method that conforms to it does, in fact, provide accurate results for real target system. There is a danger that the mere existence of rigorous benchmarks encourages belief in the general validity of a simulation method, once it reproduces them, even when the benchmarks – as can happen with rigorous results – have no connection to a model's actual performance in specific empirical contexts.

What, then, is the proper role of rigorous results and how can they help alleviate the threat of the simulationist's regress? In order to see how, it is important to realize that rigorous results have no analogue at the level of experimental data. They are genuinely new contributions at the level of mathematical models, and are often quite specific to certain classes of models. They also cannot be 'read off', so to speak, from fundamental theory: they do not merely re-state fundamental theoretical assumptions. This partial independence from both theoretical assumptions and empirical data introduces a genuinely new layer of assessment: simulations of mathematical models can be evaluated in terms of whether or not they conform to relevant model-specific rigorous results. The qualifier 'relevant' is important, of course. Given that it may or may not be possible to give rigorous results an empirically meaningful interpretation, some rigorous results will inevitably be more important than others. For example, in one model a conserved quantity may indicate the preservation of particle number, while in another model a constant number may refer to an 'unphysical' (i.e., not even in principle realizable) limiting case. There may be entirely legitimate trade-offs – that is, situations in which an investigator might tolerate the violation of a rigorous result by a given simulation method, if there are principled reasons for regarding the rigorous result as of secondary importance. As an example, consider the so-called Mermin-Wagner theorem, according to which there is no long-range order (and hence no associated phase transition, such as spontaneous magnetization) at finite (non-zero) temperatuures in one- or two-dimensional lattice systems. The Mermin-Wagner theorem has been rigorously shown to hold for a number of mathematical many-body models, including many (such as the Heisenberg, Hubbard, and Kondo-lattice models) that are currently used to describe, for example, spatially extended (three-dimensional) ferromagnetic systems. Interestingly, computer simulations of these models for low dimensional (1D or 2D) systems sometimes do predict a phase transition to an ordered state. That is, such simulations strictly *violate* rigorous results that have been established mathematically for the underlying models. What should a simulationist do in such a situation? On the one hand, low-dimensional order (e.g., magnetic layers) are an experimentally established phenomenon, and there may be independent theoretical grounds to believe that those mathematical models that successfully describe the three-dimensional case should also extend to low-dimensional samples of the same substance. On the other hand, the simulation results directly contradict a mathematically rigorous result. Interestingly, there is no agreement among practitioners in the field about how to proceed. In the example at hand, some scientists argue that the simulation 'approaches the physical reality (magnetic order in thin films does really exist) better than the Hubbard model in its exact form', whereas others insist that if a simulation 'predicts the occurrence of spontaneous magnetization in one and two dimensions as well as in three dimensions […], the validity of these predictions in three dimensions should clearly be investigated more fully'.[10] As this example indicates, the existence of rigorous results merely creates a *presumption* against their violation. But rigorous results do introduce a new *kind* of standard, by which simulations of models can be judged in terms of their relative performance, one that is largely independent from empirical performance and theoretical expectations.

The standard response from robustness, discussed in Section 6, conceived of robustness explicitly as the concurrence of measurement techniques or, more generally, techniques of data-production. On such an interpretation, computer simulation would seem to have little to contribute, since simulated data is hardly independent enough to add significantly to robustness: as Humphreys rightly notes, 'when simulated data rather than real data are fed into the simulation, the prospects for informing us about the world are minimal' (Humphreys 2004: 134-135). An exclusive focus on measurement techniques, however, would likewise construe the notion of robustness too narrowly. It is more fruitful, I believe, to think of robustness as, in Bill Wimsatt's terms, 'a family of criteria and procedures'. Among the heuristic principles that should guide robustness analysis, Wimsatt goes beyond independence of 'derivation, identification, or measurement processes' to explicitly include three further principles, namely a) the search for factors 'that are *invariant* over or *identical* in the conclusions or results of these processes', b) the '*scope*' and '*conditions*' of

[10] For sources and a brief discussion see (Gelfert 2005: 733-737).

such invariances, and c) the analysis and explanation of ‘any relevant *failures of invariances*’ (Wimsatt 1981: 126). Of the four classes of considerations that Wimsatt identifies, the independence and diversity of measurement techniques makes up only one; the remaining three deal with issues of invariance or its failure. One way in which invariances manifest themselves in mathematical models are precisely the kind of rigorous results discussed earlier, especially where these relate to conserved quantities, symmetries (as well as symmetry-breaking), scale-invariance and so forth. However, not all rigorous results concern ‘global’ constraints, such as conservation laws or symmetries of a system. As the example of the Mermin-Wagner theorem shows, rigorous results also make predictions concerning the presence or absence of certain phenomena (such as spontaneous order) in specific systems. Other examples of rigorous results include mappings of different classes of models onto one another in specific limiting cases (as in the Mott insulator scenario), upper and lower bounds on various quantities (or on features such as the correlation length) in a system, critical exponents in the vicinity of phase transitions, or the topological relationships between orbits (e.g., in systems governed by chaotic dynamics). Given this rich panoply of rigorous results and relations, it is therefore eminently sensible, when one is simulating mathematical models, to test one’s simulation methods against rigorous results – whether or not these can be given an empirical interpretation – and to come to a considered judgment whether possible violations, or failures ofcertain rigorous results, are significant or not. Rigorous results do not wear their significance on their sleeve, as it were, but they do provide an additional layer of assessment – one that is specific to the practice of simulating mathematical models and has no analogue in traditional experimental practice.

9. Conclusion: Deflating the ‘simulationist’s regress’

Where does this leave the initial worry that simulation studies may be subject to a *simulationist’s regress* that is analogous to, but conceptually distinct from, the experimenter’s regress and that may very well be more difficult to resolve in practice, given that independent data is so much harder to come by than in the case of traditional experimentation? On the account I have sketched, this worry appears somewhat exaggerated. It overstates the ‘detachedness’ of computer simulation from conditions of real inquiry, and it creates the false impression that the robustness of an inquiry is purely a matter of what kind of data it draws on, and which techniques it employs to produce such data. Once the original notion of robustness as a family of criteria and procedures is restored, it becomes apparent that the practice of simulating mathematical models is misdescribed if one thinks of it merely in terms of the generation of simulation results and their external validity. Over and above the comparison between empirical and simulated data, there exist a range of procedures of assessing whether a simulation 'respects' the fundamental features of a model; rigorous results are one such class of examples. Their status as features of mathematical models themselves – i.e., as genuinely new contributions *at the level of models,* lacking an analogue in traditional experimentation – allows them to function as independent internal standards for the assessment of simulation methods.

Worries about data-technique circles deserve close attention and they can indeed become problematic in the scenarios discussed earlier (e.g., when a dispute arises over the existence of causally isolated novel phenomena). But simulation does not, in and of itself, have to bear a heavier burden of proof than traditional experimentation. One might naively think that simulation is necessarily one step further removed from nature and that, therefore, it is somehow doomed to take place 'in thin air' – thus making it prone to a special kind of regress, where the best or only test of a simulation is its own disputed result. But such a view ignores that simulation can help itself to a range of independent internal standards, rigorous results being one example. While simulation no doubt introduces a qualitatively new step in the multi-step process of investigating nature, it also provides new criteria and ways of assessment. It thus seems appropriate to end on a deflationary note, by stating that, while simulation makes qualitatively new demands on the researcher who must exercise his judgment in assessing its results, it is no more prone to data-technique circles than traditional experimentation.

Acknowledgments

I am grateful to two anonymous referees for their constructive criticism, as well as to Paul Humphreys and Cyrille Imbert for valuable comments on an earlier draft.

Bibliography

Robert W. Batterman: "Asymptotics and the role of minimal models", *British Journal for the Philosophy of Science* 53 (2002) 21-38.

Hasok Chang, *Inventing Temperature: Measurement and Scientific Progress*, Oxford: Oxford University Press 2004.

Harry Collins: "Replication", in Sal Restivo (ed.), *Science, Technology, and Society: an Encyclopedia*, Oxford: Oxford University Press 2005, 456-458.

Harry Collins: "The experimenter's regress as philosophical sociology", *Studies in History and Philosophy of Science* 33 (2002) 149-156.

Harry Collins, *Changing Order: Replication and Induction in Scientific Practice*, London: Sage 1985.

Athel Cornish-Bowden and Robert Eisenthal: "Computer simulation as a tool for studying metabolism and drug design", in Athel Cornish-Bowden and María Luz Cárdenas, *Technological and Medical Implications of Metabolic Control Analysis*, Dordrecht: Kluwer 2000, 165-172.

Sylvia Culp: "Objectivity in experimental inquiry: breaking data-technique circles", *Philosophy of Science* 62 (1995) 430-450.

Allan Franklin, *The Neglect of Experiment*, Cambridge: Cambridge University Press.

Allan Franklin: "How to avoid the experimenters' regress", *Studies in the History and Philosophy of Science* 25 (1994) 97-121.

Allan Franklin: "Experiment in physics", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2009 Edition), online at http://plato.stanford.edu/archives/spr2009/entries/physics-experiment, accessed 01 November 2009.

Peter Galison: "Computer simulations and the trading zone", in Peter Galison and David J. Stump (eds.), *The Disunity of Science: Boundaries, Contexts, and Power*, Palo Alto: Stanford University Press 1996, 118-157.

Axel Gelfert: "Rigorous results, cross-model justification, and the transfer of empirical warrant: the case of many-body models in physics", *Synthese* 169 (2009) 497-519.

Axel Gelfert: "Mathematical rigor in physics: putting exact results in their place", *Philosophy of Science* 72 (2005) 723-738.

Francesco Guala, *The Methodology of Experimental Economics*, Cambridge: Cambridge University Press 2005.

Ian Hacking, *Representing and Intervening: Introductory Topic in the Philosophy of Natural Science*, Cambridge: Cambridge University Press 1983.

Roger W. Hockney and James W. Eastwood, *Computer Simulation Using Particles*, Boca Raton: CRC Press 1988.

Paul Humphreys: "The philosophical novelty of computer simulation methods", *Synthese* 169 (2009) 615-626.

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

Peter Imhof: “Computer simulation in the controversy over limits of growth” (Working paper, urn:nbn:de:gbv:830-opus-1726; URL: *http://www.doku.b.tu-harburg.de/volltexte/2006/172/*) Hamburg: Technical University of Hamburg-Harburg 2000, 1-19.

Daniel Kennefick: “Star crushing: theoretical practice and the theoreticians’ regress”, *Social Studies of Science* 30 (2000) 5-40.

Margaret Morrison: “Models, measurement and computer simulation: the changing face of experimentation”, *Philosophical Studies* 143 (2009) 33-57.

Margaret Morrison: “Experiments versus models: new phenomena, inference, and surprise”, *Journal of Economic Methodology* 12 (2005) 317–329.

Wendy S. Parker: “Franklin, Holmes, and the Epistemology of Computer Simulation”, *International Studies in the Philosophy of Science* 22 (2008) 165-183.

Wendy S. Parker: “Does matter really matter? Computer simulations, experiments, and materiality”, *Synthese* 169 (2009) 483-496.

Hans Radder: “Experiment”, in Sahotra Sarkar and Jessica Pfeifer (eds.), *The Philosophy of Science: An Encyclopedia*, London: Routledge 2005, 268-275.

Hans Radder, *In and About the World: Philosophical Studies of Science and Technology*, Albany: SUNY Press 1996.

Nils Röller: “*Scientia media*: simulation between cultures”, in Andrea Gleiniger and Georg Vrachliotis (eds.), *Simulation: Presentation Technique and Cognitive Method*, Basel: Birkhäuser 2008, 51-62.

David Ruelle, *Statistical Mechanics: Rigorous Results*, New York: Benjamin 1969.

Thomas P. Weissert, *The Genesis of Simulation in Dynamics: Pursuing the Fermi-Pasta-Ulam Problem*, Berlin: Springer 1997.

William C. Wimsatt: “Robustness, reliability, and overdetermination”, in Marilynn B. Brewer and Barry E. Collins (eds.), *Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell*, San Francisco: Jossey-Bass Publishers 1981, 124-163.

Eric Winsberg: “A tale of two methods”, *Synthese* 169 (2009) 575-592.

Eric Winsberg: “Sanctioning models: the epistemology of simulation”, *Science in. Context* 12 (1999) 275-292.