The Argument from Surprise

Penultimate Version, forthcoming in the Canadian Journal of Philosophy

Adrian Currie,
Cambridge (CSER & CRASSH)
ac2075@cam.ac.uk

Acknowledgements:

This paper has benefited greatly from comments by Andrew Buskell, Alice Murphy, Emily Parke, David Colaco, Arnon Levy, Mark Migotti, Marta Halina, Kirsten Walsh and Mihail Cernea. It was presented at the University of Calgary Philosophy of Biology Research Group, at the University of Washington philosophy department, and the *Simplification & Distortion as Scientific Strategy* workshop at the IRH in Bucharest – I am grateful of the feedback provided. Some of the research for this publication was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation.

Abstract

I develop an account of productive surprise as an epistemic virtue of scientific investigations which does not turn on psychology alone. On my account, a scientific investigation is potentially productively surprising when (1) results can conflict with epistemic expectations, (2) those expectations pertain to a wide set of subjects. I argue that there are two sources of such surprise in science. One source, often identified with experiments, involves bringing our theoretical ideas in contact with new empirical observations. Another, often identified with simulations, involves articulating and bringing together different parts of our knowledge. Both experiments and simulations, then, can surprise.

Keywords: Simulation; model; surprise; experiment

1. Introduction

> Machines take me by surprise with great frequency (Turing, 1950, 450).

Scientists are in the business of generating a variety of epistemic and pragmatic goods—true propositions, good explanations, veridical representations, accurate predictions, successful methods of intervention, new treatments, technologies and phenomena. By contrast, analytic philosophers of science have traditionally taken a narrower focus: on the nature and dynamics of theories, their relationship with the world and their confirmation[1]. But what matters epistemically about an investigation's results are not exhausted by the support they provide to propositions pertaining to natural systems. At least one other epistemically relevant feature is the productivity of scientific results. That is, their capacity to surprise us in fruitful ways. This is our target here.

I will develop an account of productive surprise which is, properly speaking, epistemic; that is, it shouldn't be considered a merely pragmatic virtue, nor a purely psychological feature, but is instead a quality the attainment of which constitutes genuine epistemic progress. Further, this account sheds light on the various capacities and differences between various scientific strategies. I'll probe productivity via recent discussion of experiments and simulations.

It is often thought that there is an epistemic difference between experiments and simulations; a difference favouring experiment. That is, experiments provide epistemic goods which simulations cannot. However, it has proven difficult to pin down just what this difference might be: which epistemic goods, and in virtue of what do simulations fail to provision them[2]?

[1] One well-travelled philosophical path where surprise matters is discussion of novel facts. However, as this is geared towards theory-confirmation and issues of realism it is tangential. The same is true of discussions of 'anomalies' in, for instance, Kuhn's work.

[2] There is a growing body of work comparing the epistemic status of simulations vis-à-vis experiments. Many philosophers approach the difference in ontological terms. Perhaps experimental subjects are materially continuous with their targets, while simulations and their targets are 'made of different stuff'—

One possibility is that experiments, but not simulations, are a source of ‘surprise’. That is, they can defy our expectations in ways which generates new knowledge and drives new discovery. This is the argument from surprise (see Morgan 2005, for rebuttals see Boumans 2012, Parke 2014). In contrast, I claim that both experiments and simulations can surprise, but do so on differing grounds. Experimental surprise relies on experimental *freedom*: aspects of experimental behaviour must not be too constrained. Simulations, by contrast, surprise when understood as embedded in a set of validation practices. The source of surprise, then, differs: where experiments surprise because the world behaves differently to how we expect, simulations surprise because the connections between, and consequences of, our theoretical, conceptual and empirical knowledge is often obscure and the practice of making them explicit and probing them can produce unforeseen and significant results[3],[4].

In section 2 I’ll present a case study to illustrate the relevant scientific practices. In section 3 I will examine scientific surprise, drawing particularly on Morgan, Parke and Boumans’ work. This sets us up for section 4, where I will present both a new argument from surprise and an analysis of what scientific surprise consists in. In section 5 I respond to the argument, showing how simulations can surprise after all, and in section 6 I’ll reflect on the differences between simulations and experiments.

---

and perhaps this makes a difference to the kinds of epistemic tasks they can perform (see Morgan 2002, 2003, 2005, Harre 2003, Guala 2002, 2005, Parker 2009, Winsberg 2010, Parke 2014). Others, such as Parker (2008) and Winsberg (2003, 2009) compare the two via their epistemic capacities. Although this is similar to my approach, neither discusses surprise.

[3] There are several discussions in the philosophy of modelling which I avoid here. One important question concerns the metaphysical nature of models and their relationship to the world (Levy 2012, Godfrey-Smith 2009, Weisberg 2013). Another question focuses on the potential roles models can play: building better theories (Wimsatt 1987), providing explanation (Weisberg 2007), or traction on complex systems (Mitchell 2002, 2003). A final question asks after the relationship between models and other scientific tools, such as whether models are autonomous of theory (Winsberg 2010, Morgan & Morrison 1999). I will restrict myself to this last kind of question.

[4] An area where philosophical attention has focused on the epistemic potential of simulations is in the philosophy of climate science. See, for instance, Parker 2009b, 2010, Epstein & Forber 2013, Steele & Wendl 2013, Lloyd 2009, 2010. These papers have a narrower focus than mine, and I expect much of what I say is complimentary.

In comparing experiments, simulations and other scientific techniques or strategies, philosophers often fail to distinguish between different epistemic tasks: comparisons are made as if there is some *tout court* sense in which one might be better than the other. This is too blunt: scientists have many different epistemic aims and I doubt there is anything general to say about the advantages or disadvantages of any epistemic technique divorced from those goals (Parke 2014). Accordingly, I am focusing on the capacity to surprise, and am largely ignoring other capacities such as confirmation. Undoubtedly there are places where such discussions overlap, but I leave that for later work. Epistemic progress, of course, is not limited to the production of surprising results. Scientific progress is complex and multi-faceted, and doesn't exclusively rely on any one epistemic property, practice or value (Currie forthcoming, chapter 13). Further, I'm going to be purposefully reticent about providing an explicit characterization of 'experiment' and 'simulation'. I'm not here in the business of *defining* the two practices, rather I'm ultimately interested in how different epistemic tools can generate surprise in different ways. As such, the illustrative discussion in section two and the loose characterization I provide there are suitable for my purposes.

So, I'm interested in the epistemic roles different scientific tools play. Scientists are in the business of discovering, explaining and understanding the world, and they use an array of techniques and tools to acquire the associated epistemic goods. Simulations and experiments are two of these tools. By examining the argument from surprise, I'll provide an account of what scientific surprise is, we'll also learn how simulations differ from experiments as tools, and how they succeed in generating knowledge by stymying, and motivating, scientific progress.

2. The Giant's Gait

Let's begin by examining a case involving two investigative practices. We'll take one as our representative simulation. It involves the use of a computer model to infer the gait of an extinct lineage of dinosaurs. The other is, for our purposes at least, a representative experiment. It is a dissection study which probes the relationship between gait and morphology in vertebrates.

The sauropod lineage boasted the largest terrestrial animals ever. The recently discovered *Dreadnoughtus schrani,* for instance, managed lengths of 26 meters and weights upwards of 60 ton (see Lakrovara et al 2014). Such animals present many puzzles[5], one being their gait. The mechanics of scale say that as animal size increases, some activities become exponentially more difficult[6]. How sauropods managed to shift their bulk, then, is a good question. And a difficult one: paleobiologists are working with limited remains and lack appropriate extant analogues.

Sellers et al respond to this problem by, in brief, simulating a sauropod and teaching it to walk. The strategy is to "… construct a computer simulation of sufficient biofidelity to capture the necessary mechanics of the system and to use this to test specific locomotor hypotheses" (2). Theirs is a sophisticated piece of science, and I only summarize the relevant parts here[7].

Simulation building begins with the digital capture of a reconstruction of the sauropod skeletal system, in this case *Argentinosaurus huinculensis*. This skeletal anatomy is then represented digitally (see fig 1). The skeleton is divided into segments which are treated as rigid, modular parts (think of action figures with adjustable limbs, see fig 2). Sellers et al go on to estimate the distribution of mass, and model both muscle and joints. Significant inference and idealization is involved. Rather than representing an accurate picture of sauropod muscular anatomy, they aim for *functional equivalence*:

> …it makes sense to reduce the model's complexity by using a more idealized set of muscles that represent the functional actions that are likely to be available. These

[5] See, for instance, the papers collected in Klein (2011)

[6] The effects on body size against locomotor performance depend upon whether an activity is power limited, such as jumping, or force limited, such as standing. As muscle power increases roughly proportionately to muscle mass, increases in size will be matched by increases in power. However, muscle force is (roughly) proportional to muscle area—and the ratio of muscle-area to body size decreases as size increases. The result is that some activities, such as walking, become increasingly trickier at larger sizes.

[7] Sellers et al's own account is gratifyingly clear, and I refer the reader to them.

muscles can be defined with arbitrary paths and moment arms so long as they produce equivalent actions to anatomical muscles (6).

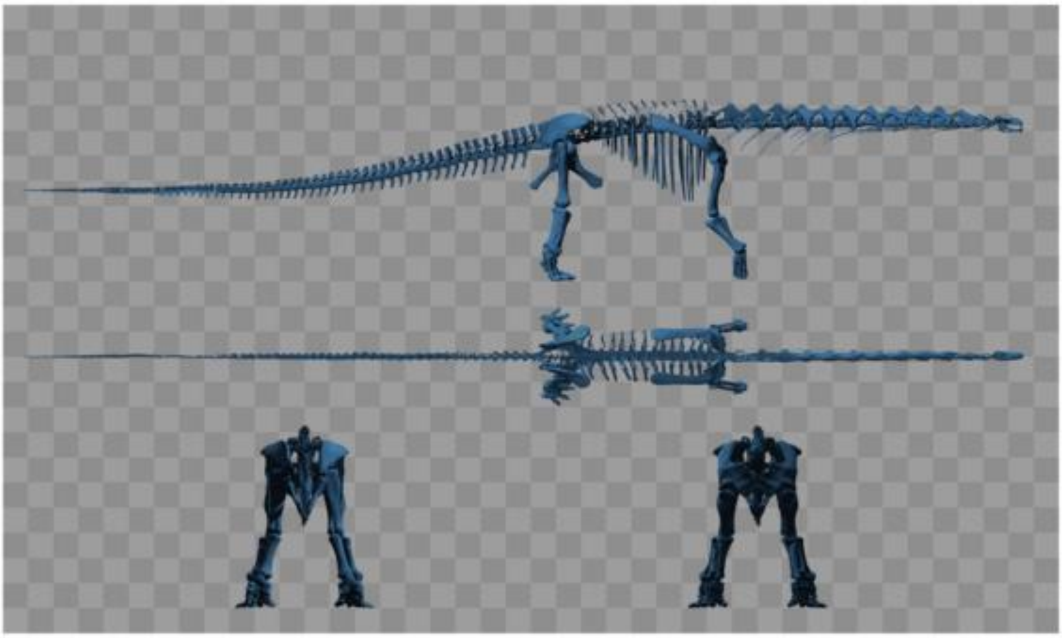

Figure 1: The digitized *Argentinosaurus* skeleton (from Sellers et al, 3, creative commons)

Finally, they estimate muscle mass. This matters because power and force are related to mass, and thus can be inferred from it (see footnote 6). In determining muscle mass, Sellers et al need to understand the relationship between gait, morphology and musculature in large cursors. With an obvious lack of living sauropods available, they instead draw on living mammals. We'll take this study as a putative experiment.

To estimate muscle mass, Sellers et al survey a group of extant cursors (quadrupeds built for running): reindeer, hare and greyhound. The data was collected by a technique which is relevantly experimental for our purposes. Dead specimens are examined, dissected, and measured in some detail in order to determine the relationship between the animal's size, facts about its lifeways, and its muscle mass and distribution. Sellers et al find that "there is a relatively consistent pattern even for quadrupeds of different sizes and locomotor specialisations" (8). On this basis they infer various muscular properties for their simulant.

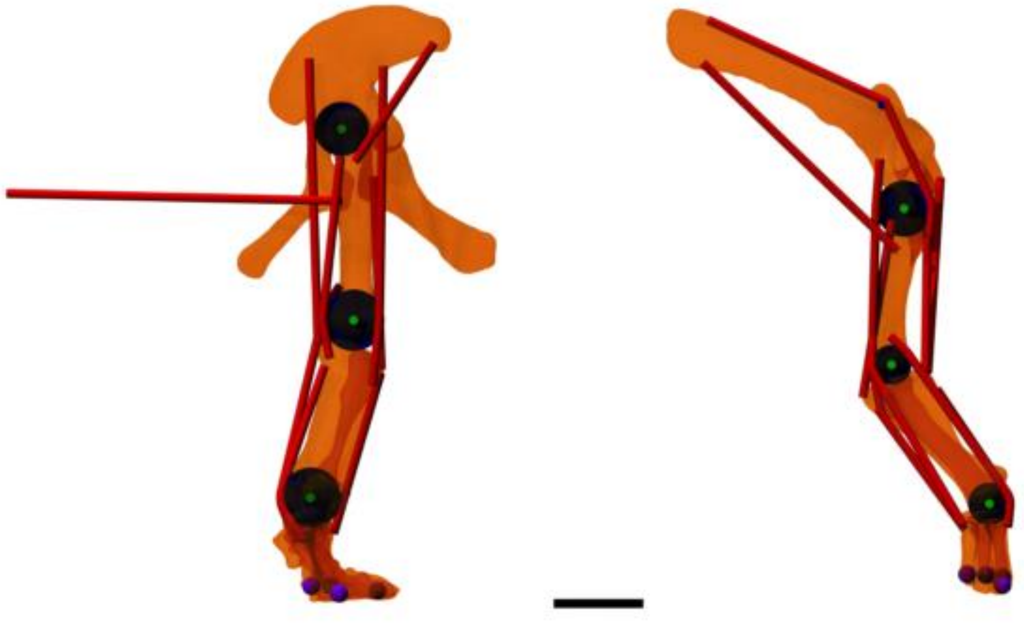

Figure 2: simulated *Argentinosaur* legs, with segments, muscle and joints (from Sellers et al, 4, creative commons)

With the simulation prepared, Sellers et al see if their simulant can walk. They run a series of simulations targeting various optima, using results to scaffold further simulations. In brief, the simulants are given an optimum target, such as maximal forward motion over a fixed time period. Various morphs, differing in their muscle activation patterns, are 'bred' and those which best achieve the target are used as the basis for the next generation.

Their results are robust. The simulant walks in a largely elephant-like way, however it could only do so "… by allowing the wrist joint to lock at a position of maximum flexion and producing a gait somewhat reminiscent of a chimpanzee knuckle walking" (13). A sauropod walking as an elephant does would catastrophically snap its forelimbs. By locking the wrist-joint, strain is distributed amongst stronger muscles. Thus, their simulation produces a unique gait, unseen in extant animals.

Importantly, Sellers et al also generate the trackway patterns such a gait would produce. At medium paces, these are remarkably similar to preserved sauropod trackways. They conclude that although the largest sauropods could happily walk,

> … it is clear that this is approaching a functional limit… Much larger terrestrial vertebrates may be possible but would probably require significant remodeling of the body shape, or significant behavioral change, to prevent joint collapse (18-19).

Their study, then, makes a prediction: if larger sauropods or other terrestrial animals are discovered, their anatomy will diverge from *Argentinosaurus.*

To summarize, Sellers et al digitize a reconstructed *Argentinosaurus*; clothe her in muscle; 'evolve' gaits by running simulations from various initial conditions. Their results converged on a "perfectly plausible" (1) elephant-like gait, made possible by locking the wrist. The model also produced trackway-patterns which matched remnants, and made predictions about future discoveries. In section 5 I shall suggest this study is surprising: it generates new, perhaps unintended, knowledge of the target, and opens fruitful avenues of empirical investigation.

Notice two practices involved in the above study. First, Sellers et al construct a virtual sauropod—their simulant—and examine its behavior under differing conditions. Second, they investigate the relationship between muscle mass and gait by dissecting a range of vertebrate cursors. The dissection studies are not paradigm experiments—they are more like generic measurements—but for our purposes are suitable. I want to highlight a property these practices share, and a way in which they come apart. Both involve the exertion of *control*: scientists systematically isolate and manipulate systems to generate results. However, the latter (the experiment) involves interacting with a representative of the class which the scientists are concerned with. The former (the simulation) involves a proxy. Rather than directly examining a specimen of the relevant type, Sellers et al investigate something they hope makes for a good representation of that type. In the dissection, we are interested in muscle distribution across vertebrate cursors, and so we examine vertebrate cursors. In the simulation, we are interested in the gait of a long-extinct lineage, and examine a digital representation.

For present purposes we can take the computational study to be an exemplar simulation: Sellers et al conduct a controlled investigation of a proxy, whose behavior is repeatedly observed under varying conditions, and is taken to be informative of their target (Parker 2009). The dissection study will serve as our experiment: Sellers et al conduct a controlled investigation of a set of representative specimens. This is not, of course, a full account of experimentation, nor of simulation. However, as we'll see, for the purpose of examining surprise, this rough and ready distinction is all that is required[8].

Both of my illustrations have idiosyncrasies, differing from some other experiments and simulations. Sellers et al, for instance, use a phenomenological model: model construction begins from empirical information, rather than from first principles. Moreover, lab based experimental investigation is often very different from dissection studies. And indeed, these differences often make for an epistemic difference (see Currie & Levy under review). But not for my purposes here: as we'll see, the dissection studies have the capacity to surprise in just the way which Morgan intends and—for the purposes of the forthcoming argument—Sellers et al's simulation does not.

3. The Argument from Surprise

"The Analytical Engine has no pretensions to *originate* anything."

Ada Lovelace (quoted in Turing 1950, 450, italics in original)

Here is a difference between Sellers et al's simulation and the dissections which informed their reconstruction. Although the latter are performed under artificial conditions, the

[8] For a fuller defence of the account implied here, see Currie & Levy (under review)

investigators are not in complete control of their results[9]. The naturally occurring properties of the dissectees, the hares, say, determine their measurements. The simulation of *Argentinosaurus*, by contrast, is not like this: although they utilized empirical information, the investigators stipulated, programmed, and constructed the simulation themselves. This difference has *prima facie* epistemic consequences. It is a tempting thought that, in some sense, simulations don't give you anything more than what you put in. The argument from surprise relies on just this kind of difference. Although Emily Parke argues that simulations can surprise, she captures the basic thought succinctly:

> While experimenters usually design at least some of their object's parts and properties, they never design all of them, and in some cases none of them, as in field experiments. A simulationist's object of study, on the other hand, is a model: she made or programmed it herself, so knows all of the relevant facts about its parts and properties. It is thought that experiments, in virtue of the nature of their objects, can thus surprise us in ways that simulations cannot (Parke 527).

Mary Morgan (2005) distinguishes between an investigation's capacity to 'surprise' and 'confound'. Mere *surprise* is simply an unexpected result—clearly simulations can provide these. However, a *confounding* result generates new research by challenging orthodoxy, generating new phenomena, opening new avenues of investigation, and so on. Following the relevant difference, I will call the latter *productive* surprise[10]. Simulations, Morgan argues, may surprise, but cannot do so productively. The incapacity to provide productive surprise underwrites an epistemic distinction between simulations and experiments.

Emily Parke (2014) and Marcel Boumans (2012) read Morgan's notion of productive surprise as involving the "researcher's epistemic states regarding results borne out in their research"

[9] In other contexts, we may want to distinguish between experiments and more 'passive' observations (Currie & Levy under review). However, this is unnecessary for the purpose of distinguishing between experiments and simulations in terms of surprise.

[10] Mark Migotti suggested the term 'productive'.

(Parke 528), that is, investigators' knowledge. We can thus distinguish between *phenomenal* surprise—the sheer *feeling* of surprise—and surprise in the sense relevant here, an occurrence which is unexpected given particular epistemic or doxastic states[11]. Naturally, how surprising a study might be in this sense will depend somewhat on the strength of these states. *Ceteris paribus,* the higher one's credence that a result will be a certain value, the more surprising it will be when it comes out another[12]. Morgan is interested in this epistemic sense of surprise: while experiments can upset our epistemic states in a way which leads to new research and knowledge, simulations (or so the story goes) either cannot do so, or do so badly. This is because we already know what has gone into the simulation's construction.

In response, both Parke and Boumans point out that scientists are often ignorant of pertinent facts about their simulations. Simulations are usually constructed by multiple researchers, many of whom are in no position to understand the complex layers of code they build on. In virtue of this, simulations often generate epistemically or doxastically unexpected results. Moreover, human scientists are not logically omniscient: we typically do not know all of the consequences which arise from a set of initial conditions, even in the relatively constrained, simple circumstances of computer simulations. Researchers, then, will often have expectations—doxastic and epistemic states—which are foiled by the simulation's actual behavior.

Moreover, "[d]ifferences in researcher's epistemic states, alone, seem like the wrong grounds for tracking a distinction between experiment and simulation" (Parke 258). Parke does not deny that some epistemic features are context-sensitive. However, such distinctions shouldn't be driven *only* by what researchers know. Presumably it is not our epistemic states which matter, but what the simulation can tell us about the worldly target of our investigation. Parke and Boumans provide putative examples of simulations which apparently are sources of

[11] Thanks to Arnon Levy for help with notions of surprise.

[12] As such, we might want to put some restrictions on *how* surprising a result ought to be, but I leave that discussion for future work.

productive surprise. Both responses are similar insofar as they take Morgan's position to rely on a notion of surprise which is tied to individual epistemic states. As Boumans says,

> [according to Morgan] Models (can) only surprise because unexpected outcomes can be traced back and re-examined by theory. An experiment (can) confound because of a larger extent of ignorance: we may have a false or incomplete theory. Parts of the world are still not discovered and so new (confounding) phenomena may appear in an experiment (Boumans 2012, 328).

Both Boumans and Parke argue that as we are ignorant of experiments, so we are similarly ignorant of models, and thus Morgan's appeal to surprise fails. For all we know, Sellers et al lacked epistemic access to their simulation, just as they did to the inner workings of vertebrate muscle mass. In the next section, I argue this misses the power of Morgan's position. I develop a notion of productive surprise which doesn't rely problematically on researcher's epistemic states. This new notion underwrites a new argument which is not deflected by appeal to ignorance of a simulation's inner workings.

4. A New Argument from Surprise

We have seen that Parke and Boumans took the distinction between mere and productive surprise to track scientists' knowledge of their experiment or simulation alone. In this section I develop a stronger notion of productive surprise. My account is preferable. First, it avoids Parke and Boumans' criticisms. Second, as we'll see in section 5, responding to it provides an explanation of how simulations can surprise. Third, as we'll see in section 6, it provides insight into the differences between experiments and simulations. Fourth and most importantly, this notion of surprise is a *bone-fide* epistemic good, rather than being merely pragmatic or psychological.

In her comparison between experiments and models, Morgan emphasizes experimental *freedom*: the object's behaviour is not wholly dictated by experimental design.

> [E]xperiments need to be set up with a certain degree of freedom on the part of participants so that their behaviour in the experiment is not totally determined by the theory involved, nor by the rules of the experiment (Morgan 2005, 324).

Experiments produce results, and results require explanation. Our explanation of experimental behaviour had better not only refer to facts about experimental design. Indeed, I think that explanation is at centre stage:

> … the constraints of the model's behavior are set, however opaque they may be, by the economist who built the model so that however unexpected the model outcomes, they can be traced back to, and re-examined in terms of, the model (Ibid, 325).

Here is, I think, the strongest version of Morgan's argument. If we were to construe what needs explaining narrowly—restricting ourselves to the output of a simulation's equations, software, and initial values—any result the simulation provides can be explained by appealing to the simulation's features and generative capacity (this narrow reading will be questioned in the next section). That is, in principle at least the simulation's output can be explained in terms of the simulation's design and implementation. Sellers et al's simulation's behavior is wholly determined by the initial conditions, the design of skeletal arrangements and muscle anatomy, software programming, and the relevant hardware. Facts about the simulation's construction and programming exhaust explanations of the simulation's results. By contrast, facts about experimental design do not exhaust what needs explaining about experimental results. For instance, that hare muscle mass distribution follows the same pattern as that of reindeer depends in part on facts *about hare*, not simply the hare we happen to have dissected. Although both simulations and experiments can surprise us in virtue of generating unexpected results, to explain such results in an experiment requires the re-examination, reassessment and sometimes alteration of ideas pertaining to much more than the experimental context; while in the simulation case only features regarding the computational events are necessary[13].

Let's get somewhat more precise.

We can distinguish between the immediate objects which scientists interact with in their studies, and their ultimate epistemic aim. Although this distinction is fluid and shifts with scientific goals it is common and useful to draw apart *objects* of study and *targets* of enquiry (see Winsburg 2009, Parke 2014). The *object* is whatever generates an investigation's data. It is what scientists observe or intervene on. Sellers et al's simulant is their object of study. Another group of objects are the hares which were dissected and measured. An investigation's *target* is what the scientists take the data to be revelatory of—it is what we are ultimately interested in. Sellers et al's target is sauropod gait. The target of the dissection studies are patterns of muscle distribution across terrestrial vertebrates. A scientist intervenes on or observes an object in order to generate data relevant to the target. Basically, my notion of productive surprise turns on whether an object's behavior also pertains to the target.

Call the collection of models, theories and narratives pertaining to an object of study or target of enquiry the *explanatory resources* of that domain. Theories about hare anatomy, or of software functioning, are examples of explanatory resources. These resources can be drawn on to explain an investigation's results. For instance, our theories of vertebrate muscle-mass are relevant to explaining results of investigations which measure muscle mass in vertebrates; our understanding of computational software and programing are relevant to explaining the behaviour of a computer simulation. A domain's explanatory resources, then, includes anything relevant to accounting for the results of an investigation in that domain.

[13] Of course, sometimes surprising experimental results to turn out to be artifacts—due to the experimental design itself—here I am referring to *successful* experiments.

Explanatory resources have a *range*: they apply to a set of *subjects*. The subjects of a domain include objects and targets. As we saw above, one set of objects are the dissected mammals, another are the computer simulations. Targets we've seen thus far have included sauropod gait, and the relationship between cursorial motion and muscle distribution. Theories of hare anatomy most obviously range over (anatomically standard) hares, but may have a wider reach. Sellers et al argue for a theory of muscle distribution that ranges over hare, reindeer and (they hope) sauropods.

From the notion of a domain's explanatory resources, we can draw a new sense of productive surprise. The crucial idea is that some results challenge our explanatory resources in ways which require alteration or reassessment of ideas and buck our expectations across different ranges. At minimum, investigation is potentially a source of productive surprise when the explanatory resources required to explain the object's behaviour also range over the target (see figure 3). What is strikingly different about the hare dissection and Sellers et al's simulation is that, in the former case, the explanatory resources (considered narrowly) relevant to the object (a dissected hare) are also relevant to the target (hares); whereas the resources which range over the simulation (how computers operate, etc…) do not range over the target. By this argument, explanations of computer operation have nothing (relevant) to do with sauropod locomotion.

Results are productively surprising when explaining the results of the investigation give legitimate reason to change, or further investigate, the explanatory resources which range over the target. Discovering that, say, a subset of standard hares have patterned muscle distribution would affect our general theories of hare anatomy. The explanatory resources required to explain the behaviour of Sellers et al's simulation appear to lack this character.

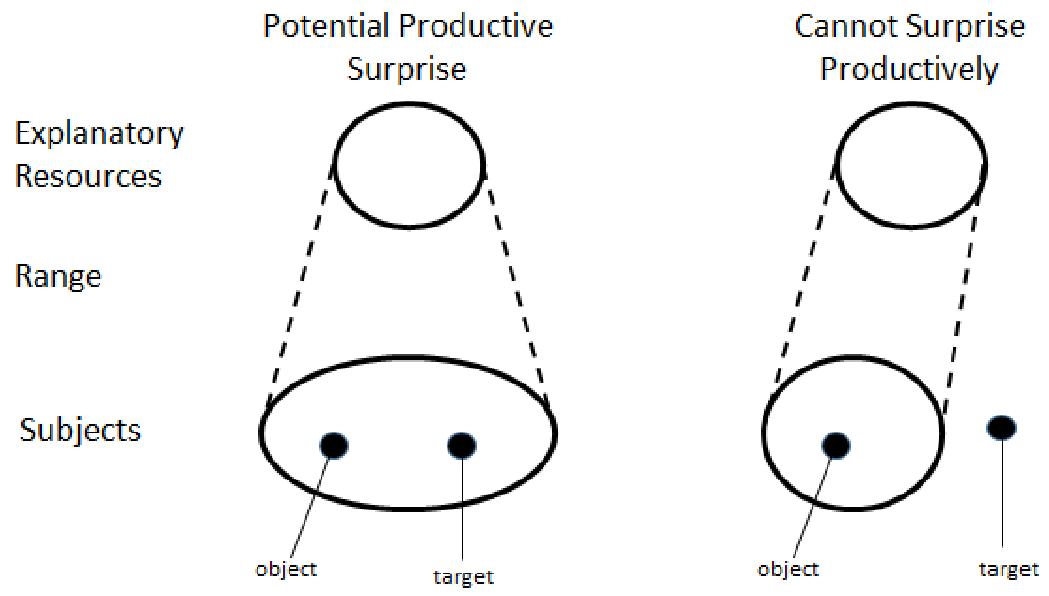

Figure 3: The difference between a potentially productive study, and one which cannot productively surprise.

An investigation surprises productively when (1) it is an epistemic surprise, that is, the object's behaviour conflicts with some doxastic or epistemic states pertaining to the object, (2) those states are *externally relevant*, that is, the same epistemic or doxastic states also pertain to the target (or a wide set of subjects), and thus lead to changes in (or at least challenge) the explanatory resources relevant to those states[14]. This is a fairly low bar for productive surprise: it is, for instance, possible to be productive without spurring new research as Morgan discusses. This low bar allows us to distinguish between different types and strengths of productive surprise. In some circumstances, for instance, we might make the following distinction. In one case, surprise occurs only insofar as an investigation's results challenge the relevant explanatory resources. In another—stronger—case, surprise includes the capacity to spur new research regarding the target. In such cases, it is the need to accommodate the surprising result into our existing knowledge which drives scientific work. This is why productive surprise has two features:

[14] I am being deliberately ambiguous between externalist and internalist readings, epistemically speaking. That is, we could read my notion of surprise internally, referring to scientific representations alone, or externally—including facts or truth. My account is amenable to a range of such readings, and I don't want (or need) to arbitrate between epistemic internalism or externalism here.

upsetting our expectations requires accommodation or updating of those expectations, a wide range means that this accommodation is potentially transformative beyond our knowledge of the object itself. As I take it the former, minimal notion, is necessary for the latter I will discuss them jointly here, only distinguishing them where necessary.

How important is the genuine feeling of surprise to my account? That is, does phenomenal surprise matter[15]? Undoubtedly, feeling surprised plays an important role in scientists' both recognizing and following up on productive, confounding results. And indeed, a fruitful and important line of enquiry would characterize how the feeling of surprise—and its role in generating scientific curiosity, perhaps—motivates and shapes scientific investigation. Here, though, we should take a different tack: the notion of surprise I have developed need not be coupled with a surprised phenomenology. It's also worth noting that productive surprise needn't rely on the epistemic status of particular individual's beliefs about the investigation at hand. Rather, it tracks the relationship between beliefs about instances and more general ideas about the behaviour of classes and the connections between them. Further, it needn't be based on the beliefs of individuals themselves: if sense can be made of a community's beliefs, or of an abstract body of knowledge, either of these could also be transformed in response to productively surprising results. On this account, surprise is a doxastic and epistemic matter, not a phenomenal one. The project of understanding phenomenal surprise in science, and its connection to the sense of surprise I've articulated here, is left for another day.

I should add a crucial point to my analysis. I have cashed out 'productive surprise' as if it pertains to a specific target. Often, however, scientific targets are ambiguous, vague, under-or-unspecified. This is particularly important for surprise. Surprising results are often unintended: they are important because of some target other than what the scientists had in mind. Moreover, experiments can be used to generate novel phenomena which drive new research (Franklin 2005,

[15] Thanks to David Colaco for pressing me on this.

Chalmers 1999, chapter 13). In such cases, productive surprise occurs when the explanatory resources relevant to a surprising result are likely to range far beyond those relevant to the object and controlled by the investigator. A strange result in a computer simulation can be traced back to facts about the computer. An odd experimental result might require resources ranging far beyond facts about the experimental setup. Thus, my appeal to 'targets' in the analysis above is merely for convenience. What truly matters for productive surprise is the potential range of the relevant explanatory resources. We can, then, define potential surprise as pertaining to a particular target, as I have above, or as pertaining to a range, like so:

> An investigation is a potential source of productive surprise *vis-à-vis* some set of subjects just when the explanatory resources pertaining to the investigation also ranges over those subjects.

The potential for an investigation to be productively surprising is an epistemic virtue: surprising results generate empirical and theoretical challenges to our knowledge which themselves generate further questions, alterations, and so on. Contrast this with the capacity for confirmation. Where the capacity to generate confirmation turns on the relationship between an observation, a hypothesis, and the background knowledge required to connect the two; the capacity to surprise turns on the relationship between the behaviour of an object, what we need to explain that behaviour, and whether that explanation pertains to further subjects. In the confirmation case, a more-or-less direct connection between object and target is desirable. In the surprise case, a wide range is desirable. Although the two are related, and in some instances overlap, they are distinct epistemic virtues. Further, both of the conditions for an investigation to be surprising in my sense—being an epistemic surprise, and being externally relevant—matter for productivity. The latter condition ensures that results can challenge ideas beyond the narrow scope of the investigation itself. The former condition motivates the challenge to our explanatory resources, and thus the subsequent productivity. Because the result upsets epistemic expectations, it drives us to update those expectations, to re-examine them; further, those changes can then lead to further investigation to bolster previous ideas or explore new ones. A surprise is truly productive when it is both externally relevant and unexpected.

With a notion of productive surprise in hand, we can turn back to the argument from surprise. The central thought is that how Sellers et al's simulation behaves is fully explained by the values fed into the machine, the software and hardware of the computer. It does not turn on *how sauropods walked.*

Compare the data from the dissection studies to the data from Sellers et al's simulation. The data produced by the former depends at least in part on naturally occurring features of the dissected animals. The results force us to re-evaluate our theories about them. For instance, I would expect an animal's means of locomotion to influence its muscle distribution. After all, whether an animal is a cursor, stotter, climber, etc… affects its physiology in relatively regular ways (Davis 1964). This initial expectation must be revised in light of the dissection data: after all, the observed collection of animals have similar muscle-distribution despite having different locomotive strategies. The explanatory resources we have about muscle distribution must be altered in light of the evidence. The study produced the data it did because of the actual muscle distribution in hare and reindeer. The simulation results, by contrast, do not appear to have this quality. The data (the resulting gait) is explained by how the computer works, how it was programmed, and so forth. Sauropods having walked differently wouldn't change that. If the epistemic connection between object and target is ruptured in this way, simulations cannot be productively surprising.

This argument is immune to Parke and Boumans' objections. First, it does not rely on scientists' doxastic and epistemic states alone, as it includes the relationship between explanatory resources relevant to target and object. And indeed the notion readily extends to communities or bodies of knowledge. Scientists' ignorance or otherwise about the mechanics of simulations are largely irrelevant. Second, although Parke and Boumans provide examples of simulations producing *tout court* epistemic surprise, they have not shown that they surprise in the sense articulated here, nor have they explained how they may do so. Although simulations can upset our epistemic expectations, according to the argument they do not upset expectations which are relevant to contexts outside of the simulation. They cannot, then, produce new phenomena, have unintended consequences, or motivate new research as experiments can. In the next section, I show why this new argument fails, and in so doing, show how simulations successfully surprise.

5. Productively Surprising Simulations

We have seen that the outputs of simulations can be explained without reference to explanatory resources which range over the investigation's target. That simulation results can be so explained matters because it purportedly shows that simulations cannot be productive. This argument fails. I'll argue first that simulations are a potential source of productive surprise, second that in some cases they are indeed productively surprising.

5.1 Surprise & Validity

The reasoning in section 4 characterizes what needs explaining far too narrowly. In the last section, I argued that the data—the outputs of simulations—can be explained by appeals restricted to resources relevant to computers. This is right, but a simulation's output is not all that needs explaining. In addition to simulation results, we must also explain *success in validation*. 'Validation', in simulation talk, is the process of ensuring that one's object is relevant to the target. Although the explanatory resources required to explain a simulation's output are restricted to facts about software and hardware, the resources required to explain the simulation's success both in aping its target, and other relevant parts of the world, are broader—or so I shall argue.

Experimentalists distinguish between 'internal' and 'external' *validity*. Briefly, internal validity establishes that the desired knowledge of the object of study is generated. Scientists ensure instruments perform properly, record the right information, and that experimental design works as intended. External validity concerns the relationship between object and target: are the experimental results projectable? Here, experimentalists test whether their investigation's relevance is not undermined by artificiality. Simulationists have parallel language. *Verification* is analogous to internal validity, while *validity* relates to external validity. Many accounts of the epistemology of simulation appeal to notions of validity and verification (see Winsburg 2010, Lloyd 2009, 2010, Parker 2008). Most of these highlight the role such practices play in what Winsberg (1999) calls a model's *sanctioning*. That is, they ask whether the simulation should be considered epistemically relevant to its target and relevant theories. They pay particular attention to the capacity of such simulations to make trustworthy predictions (particularly about aspects of the climate, see footnote 3)[16]. My argument, however, is focused on whether simulations can generate productive surprise in the sense articulated in the last section. On my view, validity considerations matter for both a model's capacity to confirm or disconfirm a hypothesis, and for its capacity to be productively surprising (although, as we'll see, being externally valid is not the source of the capacity to surprise, but a necessary condition for it).

Let's consider the validity of Sellers et al's study.

Why believe the sauropod simulation bears on sauropods? In particular, should the simulation's output—convergence on a knuckle-walking simulant, the gait necessitated by the need to distribute the animal's weight—be taken as a (at least potentially) productive result? Although it is true that, considered *directly*, the results of the study could be explained in terms

[16] It is worth pointing out that my sense of 'validity' (and, I suspect, many philosophers') is much wider than that meant by simulationists. Where scientists often refer to validity testing as a stage in an investigation which involves comparing their simulation's behaviour (or some components' behaviour) to the world, I mean any aspect which provides epistemic links between object and target.

of explanatory resources relevant to computers (what software it was running, the values entered, etc…), if we consider the investigation as a whole, many factors link the simulant to sauropods.

First, the simulant's gait is 'plausible'—it is relevantly similar (despite its uniqueness) to the locomotion of other large animals and coheres with scientists' physical intuitions. Scientists interested in morphology spend a lot of time examining gaits across a wide range of animals in different contexts. In virtue of this, they have more-or-less implicit expectations about how animals ought to move. Fitting within those expectations is reason to take simulation results seriously.

Second, the result is robust, that is, successful simulants converged on the knuckle-walk across a range of initial conditions and parameter values. This convergence suggests that results are less likely to be due to quirks of the simulation itself, and instead track regular mechanical, anatomical and physiological properties (or, at least those which were simulated!).

Third, such a gait would produce trackways similar to those left by sauropods. The gait which the simulation produced can be used to estimate the appearance of the fossils it would leave. The close qualitative match between fossilized footprints, and those modelled on the simulant's gait, provides reason to think that the originators of the fossils walked in a similar way to the simulant. Note that the trackways themselves were not involved in simulation construction or calibration, they were only brought to bear in establishing the simulation's validity.

Fourth, much of the information used to build the simulation was not stipulated, but drawn from the world. Throughout the construction process, Sellers et al stayed close to the empirical bone. They digitally captured a reconstruction which was itself based on fossil finds. Their estimates of muscle distribution and mass were based on comparisons with a wide range of extant animals. When they did idealize (for instance, in treating every joint as a ball-and-socket), they considered and tested whether these distortions would affect their results in a pernicious fashion.

Fifth, the study potentially produces predictions about future finds: anything much bigger than *Argentinosaurus* would require radically different anatomical or locomotive strategies. Given that *Argentinosaurus* is at the higher range of discovered sauropods, this result is *prima facie* promising. Moreover, this result is unintended: Sellers et al were not asking after the size limits of quadrupedal locomotion.

To some extent, then, Sellers et al's simulation was quite successful in validation. Factors involved with validation provide links both between object and target, and between the object and a range of subjects. The study accords with researchers' trained intuitions, results are robust, its outputs are consistent with independent evidence, much of the simulation was not stipulated, but was 'trained' using empirical data, and novel, unintended predictions were drawn from it. Factors like these underwrite taking the results seriously—and it is this success which underwrites their being potentially productively surprising.

What explains this success? Explaining the simulation's performance in validity testing involves, at least potentially, explanatory resources which range over the target as well as the object. Why, for instance, did the simulation produce trackways so similar to trackways left by actual sauropods? Presumably because they were produced by a relevantly similar range of motions. And our theories about that range of motions apply both to sauropods and simulants. It is in virtue of this that the simulation results license claims about extinct organisms, but further this also underwrites their being potentially surprising. Although simply explaining the *output* of the simulation does not require explanatory resources ranging over the target, once we widen our concern to include an explanation of the simulation's success in validity tests, it potentially does.

5.2 Productive Simulations

I have established that Sellers et al's work is *potentially* productive. The explanatory resources relevant to the simulation's success in validation overlap with those pertaining to sauropods. In virtue of this, it could fulfil a role analogous to that of surprising experiments: it could stump scientists, challenge their ideas, and generate further research. Is it, however, a *bone fide* case of productive surprise? To establish productive surprise, we must consider the relevant epistemic states, see whether we have good reason to modify or re-examine explanatory resources ranging over both object and target, and see whether the results drive new research.

What is surprising, epistemically speaking, about Sellers et al's results? It is not a surprise that their simulant managed to walk: after all, the critter it is modelled after surely could. What is surprising is *how* it walked: no known animal combines an elephant-like stride with knuckle-walking. So, the results went against expectations insofar as there was no expectation for that gait to emerge. It wasn't, for instance, a pre-existing hypothesis to be tested. Moreover, there were unintended results regarding maximal size in terrestrial vertebrates, and such unintended upshots are also epistemically surprising.

Which explanatory resources relevant to sauropods were challenged or altered in light of the simulation results? We have certainly learned something about our explanatory resources: that the assumptions built into the simulation can produce a walking simulant. But why think that this result should lead us to re-examine our explanatory resources which are relevant to sauropods? Again, the answer depends on the simulation study's validity. If we are to explain the simulation results as well as its success in validation, then part of the explanation will include explanatory resources which pertain to the relevant dynamics instantiated in both the simulant and the extinct sauropods. New propositions about weight and muscle distribution, and both gait speed and method will be added to our stock, and these new explanatory resources will be used when, for instance, we come across new sauropod trackways.

The results constitute productive surprise because (1) they are unexpected and (2) they promote changes to, or re-examinations of, explanatory resources pertaining to the target. So, should we go further and think that the results are productively surprising in the full-blown sense of motivating and guiding further research? To some extent, yes.

Recall my fifth example of validation: that Sellers et al's model makes a prediction about maximal sauropod size on the one hand, and gait on the other. A sauropod-like organism which weighed more than *Argentinosaurus* would need to adopt a different locomotive strategy. This leads to new questions: are there larger sauropods and, if so, do they diverge morphologically from *Argentinosaurus*, adopting different gaits? This kind of result has the hallmarks of productive surprise: new research is suggested both for fieldwork and simulation studies.

Moreover, the hypothesized gait is necessitated by the insufficient muscle mass around the wrist joint, and this solution is relevant to other questions in paleontology. Trying to understand how animals as large as dinosaurs moved is tricky, particularly considering that their apparent range of motions restricts joint muscles. One of the major challenges is determining how much muscle there is, and how they are partitioned across the joint. Even with these values, hypothesizing how larger dinosaurs—both sauropods and the predatory theropods—might have mitigated the enormous strain is difficult. As Sellers et al say, "It is particularly the case in theropod dinosaurs, with their relatively long metatarsus, that lack of sufficient ankle extensor muscle has caused problems in our earlier simulation models… " (13-14). Their work on the *Argentinosaurus* simulant provides a way of determining at least minimal muscle mass in a joint, given a range of motion, and this allows the potential testing of further hypotheses about how this restriction might be mitigated (for instance, using differently arranged tissue to increase elasticity). These techniques are eminently applicable to theropod reconstruction and, for that matter, for reconstructing large terrestrial vertebrates generally. Again, we see the productivity of Sellers et al's results.

However, the most extensive cases of productive surprise are better established in retrospect, as the productivity of a result or study depends to some extent on what research is subsequently generated. Sellers et al's study is productive: it has generated new knowledge which has the potential to drive new research. But how much new research, and how successful it might be, is an open question.

And so, the argument from surprise fails: simulations can be a source of productive surprise. Moreover, how they do so is revelatory. The epistemic properties of computer simulations require us to explain simulation-behaviour in terms of validation (and verification). And their capacity to provide productive, even unintentional, results is underwritten by this. So, simulations can surprise, but do they do so as experiments do? In the next section, I'll argue the answer is no.

6. Sources of Surprise

Does the preceding discussion tell us anything about the differences between experiments and simulations? Recall from the introduction that I'm not in the business of drawing a strict division between two classes of scientific tools or strategies: I'm inclined to think that useful divisions are context sensitive, and so in some contexts we may want to distinguish between practices such as Sellers et al's simulations and the dissection studies which they drew upon, while in others it might be best to lump them together. However, if we ask after the features in virtue of which the former is potentially productively surprising, and in virtue of which the latter is productively surprising, one difference does present itself.

A first guess about the difference turns on the requirements of validity testing: after all, I appealed to this in my argument that simulations can surprise. However, it is not the requirement for external validity which underwrites the difference here. After all, experiments must also be checked for analogous properties. In taking the results of the dissections as relevant to vertebrates generally, Sellers et al must assume (or empirically test) that the objects in question were relevantly similar to their targets. Similar epistemic issues arise in extrapolating from either kind of object[17]. Establishing validity, then, is a necessary condition for both simulations and experiments being productively surprising (as this is required to support external relevance).

Instead, I want to suggest that the difference is due to the *source* of surprise. Typical experiments—where control is applied to a subset of the class we're interested in—involve more-or-less confrontation between the world and our theoretical knowledge. As such, the world must be allowed to speak, which is to say, control ought not undermine that confrontation. The construction and running of a typical simulation, by contrast, is at base a way of filling out, making explicit, and probing our theoretical, conceptual and empirical ideas. This is a method of knowledge-generation, but not one which fundamentally involves bringing pre-existing knowledge into contact with new empirical results. Experiment and simulation often look like very similar activities, involving the construction, manipulation and examination of relatively tractable systems. However a distinction may be drawn if we characterize one as generating surprise by bringing our theoretical ideas into contact with the world, and characterize the other as generating surprise by probing, filling out, and expressing pre-existing knowledge. Such a distinction will not, I suspect, neatly track our usual categories of 'experiment' or 'simulation', and in other epistemic contexts other distinctions may be more appropriate. The dissection study, insofar as it is surprising, is such in virtue of demonstrating that our expectations about the world, of the relationship between muscle distribution and gait, say, is somehow lacking and in need of re-examination. The simulation study, insofar as it is surprising, is such in virtue of our learning where our ideas take us. Prior to Sellers et al's investigation, we had no way of realizing

[17] Although in my view establishing external validity works differently in experiments and simulations (Currie & Levy under review).

that the distinctive knuckle-walk was an open—in fact quite plausible—hypothesis regarding sauropod gait. By combining ideas about evolutionary and morphological constraints on gait, theories about optimization processes, and about sauropod physiology, the possibility hiding within our ideas presented itself. Both experiments and simulations—read in this way—are sources of genuine knowledge, and both can be productively surprising, but the surprise has different sources.

In the case of typical experiments, where our expectations of object and target behaviour are linked, we can generate productive surprise via *freedom*. As Morgan says,

> Such new behaviour patterns, ones that surprise and at first confound the profession, are only possible if experimental subjects are given the freedom to behave other than expected by the experimenter… However, if the behaviour of those taking part in the laboratory experiment is entirely constrained, then the results will be determined absolutely by the experimental design and rules. (Morgan 2005, 324-325).

Insofar as they are interested in productively surprising results, experimenters must ensure that their studies are not *too* controlled. Simulations, by contrast, do not need freedom to produce surprise. Rather, careful control allows us to bring our ideas and hypotheses together, and it is in these combinations that new knowledge arises. Again, this doesn't mean that simulations can't involve control-based surprise, or that experiments must involve it. Rather, my point is that if there is a distinction to be had *vis-à-vis* surprise, it is in these differing sources.

This has consequences for what successfully productive investigations are like. In some contexts, how much control I exert, and how surprising my results can be, are potentially in conflict (although much more needs to be said regarding the nature of that conflict). As Morgan has pointed out, navigating between the advantages of control and the need to let an experimental object behave freely is crucial to an experiment's success. In other contexts, surprise does not rely on freedom: indeed, it is somewhat unclear what 'freedom' might be for some simulations. Instead, good design is sensitive to the knowledge it represents, brings together, and interacts with. Sellers et al's simulation brought together rich knowledge of vertebrate gaits, with specific ideas about sauropods, and a theoretically informed method for inferring gait given a critter's morphology, mass and muscle distribution. How this information was treated, modelled and integrated in the simulation required careful empirical and theoretical sensitivity on the scientists' part.

The capacity to surprise productively, as I have understood it, is a scientific epistemic virtue[18]. That is, it is a good-making property of scientific investigation. For some investigation to be a potential source of productive surprise, it must not only be the case that the results have the potential to conflict with researchers' epistemic and doxastic states (be merely surprising). It must also be true that those epistemic states (and the relevant explanatory resources) range over the target of enquiry. This property is virtuous because of the epistemic goods it can produce. Results can create new phenomena, undermine old theories and hypotheses, and push investigation into unknown territory.

Bibliography

Boumans, M. (2012). Mathematics as quasi-matter to build models as instruments. In Probabilities, laws, and structures (pp. 307-318). Springer Netherlands.

Chalmers, A (1999). What is this thing called science? (3rd Edition). University of Queensland Press.

[18] This is a somewhat unusual understanding of 'epistemic virtue', and I don't mean for it to be tied to virtue epistemology (see Greco & Turri 2015) necessarily. Roughly, I consider a scientific epistemic virtue to (1) be a property of an *investigation* (as opposed to a hypothesis, evidence or agent) which (2) possession of provides epistemic goods, or makes the provision of such goods likely.

Currie, A (forthcoming). Rock, Bone & Ruin: an optimist's guide to the historical sciences. MIT Press.

Currie, A & Levy A (under review). Why Experiments Matter.

Davis, D (1964) The giant panda: a morphological study of evolutionary mechanisms. Fieldiana ZoolMem 3:1–339

Epstein, B., & Forber, P. (2013). The perils of tweaking: how to use macrodata to set parameters in complex simulation models. Synthese, 190(2), 203-218.

Franklin, L.R., (2005), Exploratory Experiments, Philosophy of Science 72: 888-899

Frigg, R. (2010). Models and fiction. Synthese, 172(2), 251-268.

Godfrey-Smith, P (2009). Models and fictions in science. Philosophical Studies 143 (1):101 - 116.

Greco, John and Turri, John, "Virtue Epistemology", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.).

Guala, F. (2005). The methodology of experimental economics. Cambridge: Cambridge University Press.

Guala, F. (2002). Models, simulations, and experiments. In Model-based reasoning (pp. 59-74). Springer US.

Klein, N (ed). (2011). Biology of the sauropod dinosaurs: Understanding the life of giants. Bloomington, IN: Indiana University Press.

Lacovara, K. J., Lamanna, M. C., Ibiricu, L. M., Poole, J. C., Schroeter, E. R., Ullmann, P. V., ... & Novas, F. E. (2014). A Gigantic, Exceptionally Complete Titanosaurian Sauropod Dinosaur from Southern Patagonia, Argentina. Scientific reports, 4.

Levy, A. (2012). Models, Fictions, and Realism: Two Packages. Philosophy of Science, 79(5), 738-748.

Lloyd, E. A. (2009, June). I—Varieties of Support and Confirmation of Climate Models. In Aristotelian Society Supplementary Volume (Vol. 83, No. 1, pp. 213-232). Blackwell Publishing Ltd.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. Philosophy of Science, 77(5), 971-984.

Mitchell, S. D. (2002). Integrative pluralism. Biology and Philosophy, 17(1), 55-70.

Mitchell, S. D. (2003). Biological complexity and integrative pluralism. Cambridge University Press.

Morgan, M (2005). Experiments versus models: New phenomena, inference and surprise. Journal of Economic Methodology 12 (2):317-329.

Morgan, M. (2002). Model Experiments and Models in Experiments. In Lorenzo Magnani and Nancy Nersessian (eds.), Model-Based Reasoning: Science, Technology, Values. New York: Kluwer, 41-58.

Morgan, M. (2003). Experiments Without Material Intervention: Model Experiments, Virtual Experiments and Virtually Experiments. In Hans Radder (ed.), The Philosophy of Scientific Experimentation. Pittsburgh: University of Pittsburgh Press, 216-235.

Odenbaugh, J. (2006). Message in the bottle: The constraints of experimentation on model building. Philosophy of Science, 73(5), 720-729.

Parke, C. (2014) Experiments, Simulations, and Epistemic Privilege. Philosophy of Science. 81 (4) 516-536

Parker, W. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. Philosophy of Science, 78(4), 579-600.

Parker, W. S. (2009a). II—Confirmation and adequacy-for-purpose in climate modelling. In Aristotelian Society Supplementary Volume (Vol. 83, No. 1, pp. 233-249). Blackwell Publishing Ltd.

Parker, W (2008). Franklin, Holmes, and the epistemology of computer simulation. International Studies in the Philosophy of Science 22 (2):165 – 183.

Sellers, W. I., Margetts, L., Coria, R. A., & Manning, P. L. (2013). March of the Titans: The Locomotor Capabilities of Sauropod Dinosaurs. PloS one, 8(10), e78733.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 433-460.

Turner, Derek. (2009) Beyond Detective Work: Empirical Testing in Paleontology. In: The Paleobiological Revolution: Essays on the Growth of Modern Paleontology Sepkoski & Ruse (ed) University of Chicago Press.

Werndl, Charlotte and Steele, Katie (2013) Climate models, calibration, and confirmation British Journal for the Philosophy of Science, 64 (3). 609-635.

Weisberg, M (2007). Three Kinds of Idealization. Journal of Philosophy 104 (12):639-659.

Winsberg, E. (2010). Science in the Age of Computer Simulation. The University of Chicago Press.

Winsberg, E. (2009). A tale of two methods. Synthese, 169(3), 575-592.

Winsberg, E (2003). Simulated experiments: Methodology for a virtual world. Philosophy of Science 70:105–25.

Winsberg, E. (1999). Sanctioning models: The epistemology of simulation. Science in context, 12(02), 275-292.