Erkenntnis (2023) 88:1447–1466
https://doi.org/10.1007/s10670-021-00410-z

ORIGINAL RESEARCH

The Value of Surprise in Science

Steven French[1] · **Alice Murphy**[2]

Received: 20 October 2020 / Accepted: 22 March 2021 / Published online: 30 July 2021

Abstract
Scientific results are often presented as 'surprising' as if that is a good thing. Is it? And if so, why? What is the value of surprise in science? Discussions of surprise in science have been limited, but surprise has been used as a way of defending the epistemic privilege of experiments over simulations. The argument is that while experiments can 'confound', simulations can merely surprise (Morgan, 2005). Our aim in this paper is to show that the discussion of surprise can be usefully extended to thought experiments and theoretical derivations. We argue that in focusing on these features of scientific practice, we can see that the surprise-confoundment distinction does not fully capture surprise in science. We set out how thought experiments and theoretical derivations can bring about surprises that can be disruptive in a productive way, and we end by exploring how this links with their future fertility.

1 Introduction

Scientific results are often presented as 'surprising', as if that is a good thing. Is it? And if so, why? What is the value of surprise in science? In addressing such questions discussions have tended to focus on one of two features of scientific practice: novel predictions and their role in the realism debate (see Hitchcock & Sober, 2004); and novel or surprising phenomena. In the former case, the surprise associated with the novelty is definitely a good thing as far as the scientific realist is concerned, indicative as it is of the 'mind independent' nature of the relevant theory. However, we shall have little to say about that here. In the latter, the surprise is valued because it suggests that, given the context, the relevant phenomenon is worthy of

✉ Alice Murphy
alicemlmurphy@gmail.com

Steven French
S.R.D.French@leeds.ac.uk

[1] School of PRHS, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK

[2] Fakultät für Philosophie, Wissenschaftstheorie Und Religionswissenschaft, Lehrstuhl Für Wissenschaftstheorie, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

further investigation. An example here would be the fogging of Becquerel's photographic plates, leading to the discovery of spontaneous radioactivity. Another would be the polarization of light by Iceland Spar, cited by Hacking, together with other surprising optical phenomena such as diffraction, dispersion and interference, in his well-known defence of the precedence of observation over theory (1983, p. 156). However, the specifics of such cases are also not our focus here, although such novel phenomena will offer a useful foil to our considerations.

Our aim in this paper is two-fold: first, we shall show that the discussion of surprise in science can be usefully extended to include two further features of scientific practice, namely novel thought experiments and theoretical derivations. We focus on these because thought experiments are also said to generate predictions and even, in a certain sense may be thought of as producing phenomena. More generally, both thought experiments and theoretical derivations can be thought of as producing 'outcomes', just as computer simulations do, yet there has been considerably less discussion of surprise in these cases. Given their central importance to science, there is obvious value in extending the discussion in this direction. Furthermore, both thought experiments and theoretical derivations offer a novel context in which to discuss surprise as they both involve the imagination and mental representations and thereby raise the interesting question, how can they or features associated with them, be surprising in that case? We shall come back to this below but this then relates directly to our second overall aim.[1]

This is to use consideration of these two scientific practices to put pressure on a well-known distinction between 'mere' surprise and 'confoundment', with the former associated with the outcomes of models and computer simulations, and the latter with novel phenomena (Morgan, 2005). The distinction is explicated in the following terms: a phenomenon is confounding, rather than 'merely' surprising, if it is 'both surprising and unexplainable within the given realm of theory' (2005, p. 324). Likewise, Ritson emphasizes the disruptive nature of surprising results and states that 'the kinds of novelty framed as most valuable are those that violate expectations and are difficult to incorporate into existing structures of knowledge' (2020, p. 1).

The outcome of a computer simulation, say, is argued to be only 'merely' surprising because it is ultimately explicable via the theories in terms of which the simulation was constructed. Any surprise in that case must presumably be due to the scientist's cognitive limitations when it comes to following the steps of the simulation, which may of course be complex. The core idea was captured by Wittgenstein who dismissed the value of surprise in deductive contexts:

> "The demonstration has a surprising result!"--If you are surprised, then you have not understood it yet. For surprise is not legitimate here, as it is with the issue of an experiment. There--I should like to say--it is permissible to yield to its charm; but not when the surprise comes to you at the end of a chain of inference. For here it is only a sign that unclarity or some misunderstanding still reigns' (1978, 111).

[1] We'd like to thank one of the referees for pressing us to be clearer on these issues.

Here, surprise arises because of people's epistemic limitations; 'a proof is too long to keep all its steps in mind, so something is lost from purview' (Simons, unpublished; see also French & Vickers, 2011; French, 2020). In such cases, then, the value of surprise is considered less than that of confoundment.

However, we shall argue that thought experiments and theoretical derivations may also be disruptive of expectations and be difficult to incorporate into existing structures of knowledge. Following Currie (2018), we shall call this sense of surprise 'productive surprise' and we suggest that it is more general than 'confoundment' which may be retained for surprise associated with novel phenomena.

We shall begin by outlining Morgan's arguments regarding the epistemic privileging of experiments over computer simulations and shall consider them in the context of thought experiments via two major approaches due to Brown and Norton. We demonstrate that thought experiments can surprise in a fruitful way, and that this cannot be straightforwardly dismissed as 'mere' surprise. This then leads us to the consideration of the nature and role of surprise in a broader theoretical context which we explore through the example of Einstein's derivation of $E = mc^2$. Here we shall draw on Morgan's claim that a result is confounding if it is inexplicable within a 'given realm of theory' and shall argue that a lot hinges on what counts as the 'given realm of theory'. Again, with a suitable choice of that realm, we shall argue that Einstein's result should be regarded as surprising in a productive sense and we shall conclude by indicating how such surprises can be understood as indicative of a certain 'fertility' possessed by the theory concerned and are valuable in that respect.

2 Mere Surprise and Confoundment

The use of computer simulations to study a range of complex phenomena is widespread throughout the sciences. In philosophy of science, much of the discussion has concerned how they compare with ordinary 'physical' experiments. Computer simulations have been referred to as virtual experiments, experiments in silico, or experiments without materiality. And some have claimed that 'Simulation modelling is just another form of experimentation' (Norton & Suppe, 2001, p. 92).[2] But their status as genuinely experimental has been contested as they do not intervene in the natural world and instead, it has been claimed, study 'hypothetical worlds' (Lenhard, 2018). One way in which the relation between these two practices has been explored is through Morgan's (2005) distinction between mere surprise and confoundment, originally presented via the comparison between modelling and experiment in

[2] Arcangeli has argued against what she sees as the pervasive 'bias' for the epistemological superiority of 'real' experiments, as compared with thought experiments and 'numerical' experiments (Arcangeli, 2018). Boyd has also argued that what matters for the epistemic utility of empirical results is their provenance (Boyd, 2018). With auxiliary information about data generation processes taken into account a notion of 'enriched evidence' can be elaborated that encompasses simulations. In this regard we might also mention Dardashti et al. (2017) who argued that 'analogue simulations' may play a confirmatory role in astrophysics, for example. Thanks again to one of the referees for reminding us of this further literature.

economics. Boumans (2012), Parke (2014), Currie (2018) and Beisbart (2018) have extended the discussion to computer simulations and their use across science.[3]

Although both simulations and experiments can achieve mere surprise, only the latter, Morgan argues, can achieve confoundment. This is articulated in terms of the key differences between the objects of study in experiments compared to those in computer simulations. Thus, Morgan links the surprise argument to a claim about the *materiality* of the former. Both simulations and experiments involve studying a system that "stands in" for the system that the scientist is ultimately interested in. But for Morgan, there is a core ontological difference; the object in an experiment *replicates* part of the world it stands for (albeit in a way that is simpler to manipulate), whereas the object of study in a simulation only *represents* the world outside of the simulation.[4]

This ontological difference then underpins that between confoundment and 'mere' surprise via the issue of *control*: As physical experiments are said to capture or reproduce parts of the natural world, the object in an experiment is a version of the object in nature. This means scientists are not in complete control of the experiment's results. Whereas in a computer simulation, scientists are studying something artificial that they programmed themselves and over which they, ultimately, retain control.

To see this difference, consider surprise in simulations. Scientists are often ignorant about certain features of their simulations and even if they know everything about the starting assumption of their models and the rules for how the system will change over time, these can be very complex, and they will not know all the consequences of the conditions that they started with. As Morgan highlights, finding out what follows from the initial conditions is the goal of running the simulation, and sometimes what follows can be unexpected. However, she states: 'The constraints on the model's behaviour are set, however opaque they may be, by the scientist 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' (Morgan, 2005, p. 325). Thus, a simulation's result can be fully explained by its design and implementation, incorporating the relevant theoretical presuppositions. As a consequence, it cannot confound.

On the other hand, when it comes to physical experiments the behaviour of the object under investigation is not completely controlled by the design of the experiment, and so genuinely new phenomena can emerge:

[3] A question to consider would be whether Morgan's claims should be taken as specific to economics, because of the complexity of people and their behaviour. Here, however, we follow others in generalizing Morgan's claim: we lose some general openness to nature when we simulate.

[4] For Morgan, this alone has epistemological implications: 'we are more justified in claiming to learn something about the world from experiment because the world and experiment share the same stuff' (2005, p. 323,). There are many issues with the materiality argument, including problems establishing what "materially similar" actually consists in Parke (2014) and some have suggested it is relevant similarity, not material similarity, that is important (Parker, 2009). Here, the materiality argument and the surprise argument will be treated separately (as in Parke, 2014). We take it that what is relevant to the surprise claim is that in an experiment, we are studying a material system (not that we are studying a system that is materially similar). This would suggest that analogue simulations could also confound.

> 'Such new behaviour patterns, ones that surprise and at first confound the profession, are only possible if experiments are set up with a certain degree of freedom… [so that its] behaviour is not totally determined by the theory involved, nor by the rules of the experiment (Morgan, 2005, p. 324).

There is, then, this important condition of "no over-control" in the case of experiments that have the potential to confound rather than merely surprise. In conducting a physical experiment, a scientist sets out to discover how a system will respond to an intervention. But if the system is over-controlled, then the system will not be able to react in this way. Instead, its behaviour is dictated by the set up and 'nature doesn't have anything to say' (Beisbart 2018, p. 187). This is in contrast to the example of Becquerel and the fogged photo plates.

To summarise Morgan's argument: in a computer simulation, surprising results only arise because we do not have epistemic access to all the consequences of our model before we run the simulation. But with an experiment, even within the setting of a laboratory there can be 'potential for independent action' (2005, p. 325). And when there is, we can be confronted with new phenomena that are 'unexplainable within the given realm of theory' (ibid, p. 324).

The epistemic value of confoundment lies in the fact that the relevant phenomena cry out for explanation. Confounding results are thus disruptive in a productive way: they force us to think seriously about our existing theories and motivate new research in order to find a way of accommodating the surprising results (again see also Currie, 2018; Ritson, 2020). We will now turn to the comparisons between thought experiments, 'physical' experiments and computer simulations in order to consider the extent to which the first may be surprising.[5]

3 Thought Experiments, Experiments and Computer Simulations

What is the relationship between thought experiments and 'physical' experiments? Some have taken the "experimental" aspect of thought experiments seriously, claiming that thought experiments are experiments in the same sense as lab-based experiments or are on a continuum with the latter (see Mach, 1896, p. 453). In the design of a thought experiment, certain factors are isolated, variables are controlled, and irrelevant aspects are idealised away. These variables are then manipulated and the experimenter, albeit in their imagination, "observes" what follows.

[5] There is an issue here, raised by one of the referees, as to whether surprise should be seen as a psychological notion or not; that is, is it a feeling we have, or is it an objective relation between some background commitments, outstanding problems and live methodological options? Our response is that it involves both: as we try to articulate here, the surprise associated with certain thought experiments and theoretical derivations is indicative of their significance and arises in a specific context that involves background commitments etc. but of course it manifests, in both the scientist and their audience, as a specific feeling. See Currie (2018) for an account of surprise as an "epistemic good" in science, rather than just a psychological feature.

Similarly for Brown (1986, 2007), that thought experiments take place in the "laboratory of the mind" does not entail that they are not experimental in the same sense as those that take place in the physical laboratory. He argues that thought experiments involve quasi-observation of what is essentially an abstract set up; a system is represented and then observed by the mind's eye in a way that is analogous to experiments.[6] In contrast, others have drawn a sharp line between thought experiments and 'physical' experiments. For example, for Norton (1991), thought experiments are just arguments. As they work by inferences and do not involve interacting with, manipulating nor observing the natural world, any similarities with 'physical' experiments are superficial. We shall return to both these accounts below.

There is debate, then, around the relationship between thought experiments and experiments, much of which is centred around the question whether the former should be classed with the latter or held as distinct. This makes computer simulations a helpful point of comparison when thinking about the epistemology of thought experiments given that (as we saw) there is also debate regarding whether they can be experimental in some sense. In light of these comparisons, we can now think about Morgan's surprise-confoundment distinction in the context of thought experiments. Here we are less interested in the identity question—are thought experiments or computer simulations *experiments*? Instead, we shall focus on the issue of privileging experiments in virtue of their capacity to confound rather than 'merely' surprise.[7]

4 Surprise in Thought Experiments

What does Morgan's distinction between surprise and confoundment mean for thought experiments? On one hand, we clearly know of examples of thought experiments that have produced unexpected and significant outcomes. Take Einstein's chasing a beam of light example which exposes the surprising tensions between Newtonian mechanics and Maxwell's equations. On the other hand, thought experiments, like computer simulations, do not involve interaction with the world. So should the surprise that arises from thought experiments be dismissed as a less valuable kind as Morgan suggests in the case of computer simulations? We shall show that, depending on the account of what a thought experiment is, there are alternative views as to how they can surprise, and whether they can confound. We shall first examine the issue from the perspective of Brown's platonist view, before turning to Norton's account. We then suggest an alternative position which attends to the role

[6] More specifically, this—the phenomenon, i.e. what is observed in the experiment—is what Brown would label the "narrow" sense of a (thought) experiment. Experiments in the broad sense 'includes the whole thing from theory and background assumptions to the final result' (2007, 158).

[7] See Sorensen (1992), Bokulich (2001) and Stuart (2016) for further discussions on the relations between thought experiments and experiments. While such views also have implications for the epistemic status of thought experiments, we do not discuss them here given that they do not focus on surprise.

that the imagination plays in thought experiments that demonstrates how they can bring about productive surprises in a distinctive way.

(a) Brown's View: Thought Experiments and Platonism

Brown argues that there is a set of thought experiments that provide knowledge of the world through "transcending empiricism"; they allow us access to the laws of nature that exist as relations holding between universals, such as mass, spin etc., that are taken to be platonic entities. Brown presents Galileo's famous thought experiment against Aristotle as an illustrative example. This undermines Aristotle's theory that heavier bodies fall faster than lighter ones. Galileo asks us to imagine attaching two balls together, a heavy one and a light one, and dropping them from the leaning tower of Pisa. What does Aristotle's theory predict? Both that the combined bodies will fall faster than the heavier ball on its own, as the combined object is heavier, and that the combined object will fall slower, as the lighter ball is inclined to fall slower and so, will drag the heavier body back. From this, Galileo proposes a new theory; all objects made of the same material fall at the same speed.

Brown states that here, 'we have a transition from one theory to another which is quite remarkable. There has been *no* new empirical evidence. The old theory was rationally believed before the thought experiment, but was shown to be absurd by it. The thought experiment established rational belief in a new theory' (1986, p. 10). For Brown, this is a priori knowledge; the belief in Galileo's theory is not based on new empirical data and importantly, neither is it logically derivable from old data (we shall return to this below when we discuss Norton's view).

We have already seen that Brown takes the analogy between thought experiments and physical experiments seriously. And just as the latter may confound us, so may the former on this view since this class of thought experiments may produce results that cannot be traced back to or explained in terms of the initial conditions of the thought experiment, and these results may be inexplicable in terms of the 'given theory'. Thus, for Brown, the insights we gain from platonic thought experiments are not simply a matter of 'seeing old empirical data in a new way' (ibid., p. 11) but rather, involve genuine *discovery.* Here, then, we see how thought experiments may *confound*, at least on a 'platonic' interpretation. Of course, that interpretation comes with a certain ontological cost and one might prefer to avoid that by adopting a more minimalist approach to which we shall now turn.

(b) Norton's View: Thought Experiments are Arguments

This alternative view takes thought experiments to be arguments. In answering the question of how they can have novel empirical import Norton claims that there is 'only one non-controversial source from which this information can come: it is elicited from information we already have by an identifiable argument… The alternative to this view is to suppose that thought experiments provide some new and even mysterious route to knowledge of the physical world' (1991, p. 129).

Norton's view may be separated into two claims. The first is a reconstruction thesis: The epistemic power of a thought experiment is that of its reconstructed

argument form. The second claim is about the performance of a thought experiment: the conduct of a thought experiment just is that of an argument.

Revisiting Galileo's thought experiment, it can be reconstructed as an argument (uncovering an inconsistency in Aristotle's physics) as follows:

(i) Natural speed is directly proportional to weight
(ii) Weight is additive
(iii) Natural speed is mediative

From (ii) and (iii), we get the negation of (i).[8]

Beisbart and Norton (2012) and Beisbart (2012) claim that computer simulations are also arguments. The thought is that computer simulations raise a parallel issue to the above question: how do they provide knowledge about a real-world target without any observation of that target? Their answer is that thought experiments and computer simulations provide knowledge in the same way: we build what we know into their construction, that is, the description of the thought experiment or the assumptions of the computer simulation, and this knowledge is then transformed through a logical process. Thus, computer simulations can also be reconstructed into arguments, and their epistemic force is not thereby lost. And further, that 'the reconstructing argument is executed when a computer simulation is carried out' (Beisbart, 2012, pp. 419–420). We shall not consider further this view here but we will come back to some of the worries of the argument view when applied to thought experiments.

Now, on this view of thought experiments, do they 'merely' surprise or can they confound? Beisbart and Norton do not deny that we gain new knowledge from thought experiments (and computer simulations) as 'the results inferred were not known prior to investigations' (2012, p. 409). However, they draw a distinction between 'discovery', as in the case of physical experiments and 'inferring' as in these cases, where thought experiments can be articulated in terms of inferences drawn from what is implicit. In Galileo's thought experiment, the contradiction in Aristotelian physics was already, in some sense "there"; the thought experiment *qua* argument simply exposed it.

However, reductio arguments such as this are pragmatically awkward in that the reader is invited to assume that which is subsequently shown to be false. If we take this initial assumption or premise, that is, (i) in the above reconstruction, as the 'given theory' in the characterisation of confoundment, then of course the conclusion, that the 'given theory' is false, cannot be explained in terms of that very theory (at least not on most accounts of explanation). However, if the 'given theory' is expanded to include the argument as a whole, then clearly the conclusion is explicable—we've just given an argument for it! In terms of this 'argumentative'

[8] Norton's reconstruction is limited to the "destructive" part of Galileo's thought experiment and does not include the step that is central to Brown's platonism view, i.e. the introduction of the new theory that all bodies fall at the same speed. This is because, Norton argues, the move involves a problematic assumption, namely that the 'speed of fall of bodies depends only on their weights' (1996, 342).

characterisation, then, thought experiments such as Galileo's may surprise but they do not confound.

Here we recall Wittgenstein's dismissal of the value of surprise in deductive contexts on the grounds that the cause of the surprise has to do with scientists' cognitive limitations. If we were to follow this line, along with Norton's presentation of thought experiments as arguments, then there might seem to be little of any interest to say about surprise in this context.[9] However, it is important to note that Norton's reconstructions are not limited to deductive arguments; they can also include inductive steps, as in the example of Einstein's elevator[10]; 'the case is typical and will hold for all observable phenomena' (1991, p. 137).[11] And so, Norton's view of thought experiments allow for steps that are ampliative; they go beyond what is stated in the premises.[12] The same holds too for his and Beisbert's account of computer simulations, these can also transform the assumptions in the model in a way that preserves the probability of truth (2012, p. 411).

Nevertheless, on Norton and Beisbart's view, the information we gain through deductive and inductive inferences does not constitute genuine discovery as in the case of experiments (2012, 409). And Beisbart (2012, 2018) has explicitly endorsed Morgan's account when discussing the epistemic status of simulations, offering the example of the Michaelson–Morley experiment (1887) that undermined the view that the earth has a non-zero velocity with respect to the ether. As Beisbart argues, this experiment 'has a complicated set-up, and a number of assumptions are needed to interpret its data as having implications about the ether. But this does not imply what the result of the experiment is'. If instead, a simulation was used, it would

[9] French and Vickers (2011) introduce this to undermine Popper's view that surprise motivates a realist view of theories. Again, we come back to connections between surprise in theories and thought experiments in Sect. 4.

[10] Einstein imagined someone performing experiments, such as throwing a ball and observing its trajectory, in an elevator that was accelerating upwards. He realised that the observations made would be exactly as if the elevator were in a gravitational field, concluding that the principle of relativity should be extended to include accelerating frames of reference. This thought experiment thus played a fundamental role in the development of the General Theory of Relativity. What it brings into the light is the equivalence between inertial and gravitational mass which had been noted by Newton but which Einstein elevated into a fundamental principle.

[11] Norton's notion of logical reasoning in thought experiments has expanded over the years to include steps beyond deduction and induction to informal inferences and reasoning from analogy. This has been taken to render the argument account 'vacuously true' (Brendel, 2018, p. 287).

[12] This ampliative feature may suggest that induction can lead to discoveries or generate surprises beyond 'mere surprise', and it would be interesting to think more about this in the context of an argument view of thought experiments. However, consider enumerative induction for example: after observing many white swans under various conditions, it hardly seems surprising that the next swan I see is white. Perhaps then it is the universal generalisation 'All swans are white' that is meant to come as a surprise, but again given that this would have been formulated after observing numerous swans under varied conditions, that seems implausible. What might seem a surprise is if the generalisation holds, unfalsified, even when the field is greatly enlarged (to include Australia, say) but then one might expect that feeling of surprise to dissipate as scientists offer an explanation as to why the generalisation has to hold, for fundamental biological reasons. Again, we are grateful to a referee for raising this issue.

not have confounded as there would be an assumption regarding the earth's velocity with respect to the ether in the simulation's programming (Beisbart, 2018, 12).[13]

Despite this, Currie (2018) and Parke (2014) give examples of simulations that produce results that go against expectations and 'promote changes to, or re-examinations of, explanatory resources pertaining to the target' (Currie, 2018, p. 654).[14] We have indicated above how the issue of whether thought experiments confound, rather than merely surprise, may depend on how they are characterised. However, what is crucial is that even when presented in the form of an argument, they can be disruptive in the sense of forcing us to re-evaluate our existing theories. Indeed, for many such thought experiments this is their principal role and it is obviously the case in the reconstruction of Galileo's thought experiment, where, although there are no new empirical discoveries being made, the scenario we are asked to imagine exposes a contradiction in Aristotelian physics and subsequently prompts the development of a new theory.[15] Taken together, these conclusions put pressure on Morgan's claim that the different sources of surprise impact the epistemic status of the feature under consideration.

Having said that, we agree that there is a difference between thought experiments and computer simulations on the one hand, and experiments on the other, in that the surprise arises in a different way. So, to pursue the comparison further, we recall that physical experiments can result in new empirical results that may force us to revise our theoretical knowledge. Simulations differ in that designing and running a simulation is a way 'of filling out, making explicit, and probing our theoretical, conceptual and empirical ideas' (Currie, 2018, p. 656). This is still a way of generating knowledge (and can bring about productive surprises) but unlike the experiment case, it does not involve this 'contact with new empirical results' (ibid). Likewise, thought experiments probe our theoretical, conceptual and empirical ideas. However, there are important differences between thought experiments and computer simulations which illustrate how they probe this knowledge in different ways. And this has implications for how the former bring about productive surprises.

(c) Thought experiments and the imagination

[13] Further, we can consider cases of thought experiments that may bring about mere surprise (in the sense of an unexpected consequence) but do not confound. One example, discussed by Bokulich (2001), is the rockets and thread thought experiment, which draws out a physical implication of special relativity (below we shall consider an implication of the theory that we claim can be regarded as productive in a way that suggests the distinction between 'mere' surprise and confoundment is too coarse-grained).

[14] Importantly, they each give examples of simulations which, they argue, can confound in Morgan's sense. Parke presents the example of the ABM Sugarscape which had "hidden features" that were revealed in the simulation (2014, 531). Currie outlines a simulation of sauropods' gait. The result was unexpected and prompted the investigators to reflect on the explanatory resources of the target (2018, 654).

[15] According to Feyerabend (1975, pp. 73ff; see also Arthur, 1999, pp. 220–227) Galileo offered a new 'natural interpretation' of the phenomenon allowing him to bring the Copernican view into consonance with the facts that apparently refuted it.

In order to explore those differences, let us begin with the view that computer simulations are simply more complex thought experiments. Di Paolo et al (2000) characterise simulations as 'opaque' thought experiments, and Lenhard (2018) has argued that due to their complexity and opacity, the former are more likely to surprise than the latter.[16]

Although it may seem that simulations are more transparent in that they work by a large number of simple steps, what Lenhard means is that thought experiments 'have to meet high standards of intelligibility, because the whole process takes place in cognition' whereas in a computer simulation, 'it is the multitude of interrelated steps that can render the overall process opaque' (2018, p. 485). If we take him to mean mere surprise, as opposed to confoundment, then his claim is that we are more likely to get surprising behaviours (some of which may be productive) from computer simulations than from thought experiments, as the latter are "transparent" in a way that the former are not.[17]

However, characterising computer simulations as more complex or opaque thought experiments misses something important about the latter. Firstly, part of what is surprising about thought experiments is their simplicity. There is something surprising in Galileo's thought experiment that it had such significance in the history of science, despite being a simple imagined scenario, involving the behaviour of bodies being dropped from a tower.[18] We shall come back to this in the context of theoretical derivations.

Secondly, we can see, by attending to the role of imagination, that thought experiments can bring about surprise in a distinctive way. It just is not obviously the case that we have clear access to our imaginings and the connections between them, and hence thought experiments cannot be characterised as straightforwardly as this view presupposes.[19] Thus, returning to Galileo's thought experiment, Gendler has argued that, contra to Norton's account, it is not straightforward to conclude that Aristotelian physics is inconsistent, since it is unclear whether all the propositions in the reconstructed argument form ought to be considered part of Aristotle's theory. In particular, it has been asked why we should consider (iii) as part of the theory—that natural speed is mediative, or more specifically that 'Natural speed is a property

[16] Stuart and Nersessian (2019) also discuss the different ways in which scientists can lack access to their computer model, and they argue that visualisations can be created to reduce epistemic opacity.

[17] See Lenhard (2019) for his more detailed view on surprise in simulations.

[18] Of course, the surprise generated by such thought experiments may vary according to the relevant scientific (and perhaps more broadly, social) context. Having said that, it is often difficult to discover how surprised scientists—much less, general readers—were upon being presented with one, particularly in earlier centuries, given that this reaction was typically not recorded (for an example of a modern expression of surprise over certain features of theoretical physics, see Peierls, 1979). Again, our thanks to one of the referees for reminding us of this issue.

[19] The Wittgensteinian dismissal of surprise in deductions can be linked to his claim that the imagination cannot provide us with new information because it is subject to the will (whereas 'real' objects are not) (1980, §80). Similarly, White states 'one can't be surprised by the features of what one imagines, since one put them there' (1990, 92). Stock (2007) and Todd (2020) have offered a detailed response to such claims, and see also Kind (2018) and Egeland (2019) for discussions of how we can gain new information from the imagination.

such that if a body A has natural speed 1, and a body B has natural speed 2, the natural speed of the combined body A–B will fall between 1 and 2' (1998, p. 404). Without this assumption, the inconsistency claim is unfounded.

As a result, there are various logically possible ways out for the Aristotelian. For example, they can ask—are the bodies that are tied together one object or two? If one object, then it will fall at the speed that is proportional to the combined weight.[20] Gendler contends that the thought experiment is indispensable and cannot be reconstructed in Norton's sense without losing its demonstrative force.[21] This suggests that the imagination allows kinds of jumps that cannot be accommodated within the framework of more formal reasoning.[22] Understanding thought experiments as arguments thus fails to fully capture their potential to productively surprise, a feature that characterizes, at least in part, their role in scientific practice.[23]

Our conclusion, then, is that thought experiments open up space for a discussion of surprise that is more nuanced than a classification into either 'mere' surprise or confoundment. Certainly, the productive nature of the surprise they engender suggests that the former label is inadequate, whereas the requirement of inexplicability in terms of the 'given theory' associated with the latter clearly needs to be handled carefully. With that in mind, let us now turn to a further scientific arena in which surprise can arise, that of theoretical derivations.

5 Theoretical Surprise

Consider Einstein's derivation of $E = mc^2$ which Popper subsequently declared must have come as a surprise to him (1978, p. 162).[24] Here the surprise is not that associated with discovering that a prediction turns out to be correct; that is, it is not the kind of surprise associated with novel predictions. Rather, the surprise is associated with the theoretical derivation itself, prior to any confirmation of a theoretical

[20] Indeed, it has been argued that an Aristotelian could have chosen this option—there is no commitment at this time on this issue (see Vickers, 2013, p. 196).

[21] For Gendler, Galileo's rejection of the Aristotelian view, and the "blocking" of the Aristotelian "ways out" (when the thought experiment is presented in its non-argument form) is justified because it taps into our previously unarticulated knowledge of the world (1998, 407). In this sense, her account denies the claim that imaginings are solely constituted by the person who is imagining (which was key to Wittgenstein's scepticism) since the background beliefs that contribute to the imagining come from the imaginer's experience of the world, rather than solely from the imaginer themselves. Stock (2007) also discusses how imaginings are partly informed by beliefs about the world.

[22] It is often highlighted that the imagination has to be appropriately constrained if it is to provide insights about the world. What we emphasise here is not that the imagination is totally unconstrained when it is fruitful in science, but rather that it can allow for reasoning that is less restrictive than that in arguments or computer simulations (see also Stuart, 2020).

[23] Focusing on imagination allows us to capture this sense of surprise without committing to a platonic view of thought experiments.

[24] It has been questioned whether this is such an apposite example, given, it has been claimed, that the exact meaning of this equation is contentious. Whether or not that is the case (and we think not), this is a clearly significant result about which surprise has been expressed and which also exemplifies certain features that we wish to focus on here.

prediction, which in this case had to do with the discovery of nuclear fission. For Popper, the epistemic value of the surprise in this case seems to have been the same as that of Becquerel being surprised at his photographic plates being fogged.[25] That is, just as 'material' reality may surprise us, so can theories, leading Popper to famously locate them in his World Three, or, '… the world of intelligibles, or … the world of theories in themselves, and their logical relations …' (Popper 1972, p. 154).

If we take these 'intelligibles' as abstract entities, in some sense (see French, 2020, Ch. 5) we can draw a clear comparison with Brown's view of thought experiments, as discussed above. Indeed, Popper insists that theories have a property that *only* existing things could have: this element of *surprise*. He takes this to be a mark of the reality of something: just as physical objects surprise us as we discover more about them, so too do scientific theories.

One could maintain that this is a case of 'mere' surprise and insist, along Wittgensteinian lines, that the reasons why people are surprised by such theoretical implications lie in their cognitive limitations. In other words, if Einstein was surprised it was only because not even he was logically omniscient. However, even this does not mean that it has no epistemic virtue as we saw in the case of regarding thought experiments as arguments. Before we consider that point, it is worth noting, however, that the Wittgensteinian line appears to falter in this case, simply because Einstein's proof is famously not that long, with the entire paper running for only three pages.

Einstein begins by noting that '[t]he results of the previous investigation [namely his paper setting out the basis of Special Relativity] lead to a very interesting conclusion…', an opening sentence that may indeed indicate his surprise at the result. He then invokes Maxwell's equations, which, as he notes in a footnote, incorporate the principle of the constancy of the speed of light, and the principle of relativity and applies them to the situation in which we have an extended body emitting a pair of light pulses in opposite directions, effectively outlining another thought experiment.[26]

Einstein then considers the change in translational kinetic energy of the body as a result of emitting the light pulses. The problem is, the expression for the kinetic energy of a particle is not straightforwardly extendable to that of an extended body in relativistic physics. So, Einstein defined the kinetic energy of such a body moving with speed v in a given inertial reference frame as the difference between the energy of the body in that reference frame and its energy in an inertial reference frame in which it is at rest (see Ohanian, 2009, p. 168). With this at hand, he could then obtain an expression for the change in kinetic energy of the body when it emits the

[25] Bedessem and Ruphy (2019) give the Becquerel case as an example of 'scientific unpredictability', in the sense of 'the occurrence of unexpected results in the course of the inquiry that open up new lines of research and discoveries.' (ibid., p. 3). Of course, Becquerel was already 'primed', as it were, to make such a discovery, given his interest in phosphorescence and following the discovery of x-rays by Röntgen.

[26] Thus, we might view this case as a kind of hybrid of the sort of thought experiments examined above and a 'pure' theoretical derivation.

pulses of light in its rest frame, as observed from a moving frame. Finally, he took the low-speed approximation of the energy, by neglecting magnitudes of fourth and higher orders, and substituting that in his expression he obtained, in modern form, $E = mc^2$. Interestingly, given what was to come, he concluded with the speculation that 'It is not impossible that with bodies whose energy-content is variable to a high degree (e.g. with radium salts) the theory may be successfully put to the test' (1905, p. 3). On the Wittgensteinian approach, we would expect short, simple derivations to be unsurprising. Thus, given the brevity and apparent simplicity of the derivation, this approach cannot account for the surprise felt over Einstein's result.[27]

Why, then, would Einstein, or anyone else, have been surprised, as Popper suggests?

The question becomes even more acute once it is acknowledged that some relationship between mass and energy was well-known at the time in the context of electromagnetic radiation. The likes of Heaviside, Abraham and Lorentz, among others, all investigated how the mass of a charged object changes in an electromagnetic field, yielding the notion of 'electromagnetic mass', with Hasenöhrl deriving the expression $E = (4/3)mc^2$. Poincaré (who together with Lorentz is famously associated with the 'discovery' of Special Relativity) *did* express an attitude of surprise in this context, but associated it with the conclusion that if mass, as an 'essential property of matter' is reducible to energy in this manner, then matter itself cannot be said to exist (Poincaré, 1906).

Perhaps the answer to our question lies in the observation that Einstein's result replaced the above line of research with the relationship between $E = mc^2$ and more general principles having to do with the nature of space and time (something driven home by Minkowski's 'reformulation' of the theory). In that case the surprise is associated with the establishment of such a relationship between an already known result (broadly and granted the difference in numerical factor) and these general principles that eventually came to be appreciated as underpinning a very different view of the world.[28] The answer, then, to Popper's question is that Einstein was the first to obtain *that* relationship.

Certainly, many years later, Meitner recalled her own surprise over Einstein's result when he presented it in a talk in Salzburg in 1909, writing:

> 'At that time I did not realise the full implications of the theory of relativity and the way it would contribute to a revolutionary transformation of our concepts of time and space. In the course of this lecture he did, however, take the theory of relativity and from it derive the equation: energy = mass times the square of the velocity of light and showed that to every radiation must be attributed an inert mass. These two facts were so overwhelmingly new and surprising that, to this day, I remember the lecture very well' (Meitner, 1964, p. 4; see also Rife, 2019 and Sime, 1997, p. 39)[29]

[27] Having said that, the derivation is flawed (for an overview, see Ohanian, 2009).

[28] We are grateful to Aaron Meskin for a coffeehouse conversation on this issue.

[29] It has been suggested that what surprised Meitner were the technological implications of the result.

Meitner, of course, together with Otto Frisch, subsequently used the formula to explain nuclear fission, making good on Einstein's observation in the penultimate line of his 1905 paper as indicated above.

Granted, then, the surprise associated with Einstein's result, what is its epistemic significance, if any? Again, we have to take care when it comes to the requirement of inexplicability in the context of a given theory. If that is taken to be Special Relativity itself, then clearly the result, being derived from that theory, is not inexplicable in terms of it! However, the above historical considerations suggest that we should take the 'given theory' to be the classical 'electromagnetic worldview' of the time, with any confoundment, in Morgan's sense, associated with the establishment, in that context, of the derivation of the relationship between energy and mass—some such relationship having already been posited—from a fundamental reconceptualization of space and time.

Furthermore, as in the case of thought experiments, reflecting on the surprise associated with such theoretical derivations suggests that it should be characterized as 'productive'. In the next section we shall consider how this form of surprise might be situated within an appropriate epistemic framework.

6 Surprise and Theoretical Fertility

Consider again the example of Becquerel's discovery: not only was it disruptive, in a way that might be partially, at least, captured by the notion of 'confoundment' but it was also *fruitful*. It was disruptive in that it overturned existing accounts of radiative phenomena and, ultimately of course, it contributed to the overturning of classical physics; and it was also, clearly and relatedly, immensely fruitful, with Becquerel himself publishing seven papers on the phenomenon immediately afterwards, initiating an intense programme of research involving the Curies and many others, of course. Jumping ahead over 120 years, following her survey of scientists working at the Large Hadron Collider, Ritson concluded that 'The kinds of novelty framed as most valuable are those that violate expectations and are difficult to incorporate into the existing structures of knowledge. In such instances, disruption to the existing ontology or ways of knowing were valued' (2020, p. 2).[30]

In this case, involving the discovery of novel properties of particles, she argues that scientists cash out the value of such novel results in terms of indicating a

Footnote 29 (continued)
However, granted that she is recalling her past surprise, this does not seem plausible given that such implications were not apparent in 1909.

[30] As one of the referees has reminded us, we should be careful not to generalise too far from this case study as there may be examples of results that are surprising but are not disruptive. One such that has been suggested is the recent case of the Google DeepMind 'AlphaFold' protein folding algorithm that can apparently accurately predict many protein structures from their amino-acid sequence. However, many commentators have emphasised that this does not solve the 'protein-folding problem' insofar as no explanation is given for the structures obtained and in that respect this is not a case of theoretical derivation, which is what we're interested in (see, for example, Ball, 2020).

direction for future research: 'This appraisal that potentially theoretically unexpected results can provide future fertility helps us to begin to understand how results that contradict expectations can be valued.' (ibid., p. 7). Thus, Ritson argues, the positive appraisal of disruption is based on forward looking assessments of future fertility, or forms of heuristic appraisal. She notes, in particular, the comments of scientists who are effusive in their assessment of the fertility of a disruptive result because it would point researchers in the direction of future results that might accommodate the disruption.

Here, contradicting expectations might be understood as going beyond being inexplicable in terms of a given theory and in this sense, being disruptive is broader than confoundment. The interchangeability of mass and energy is appropriately characterised as disruptive in this sense and, as expressed by Einstein, was also fertile in that it indicated the direction of future research. This sense of 'future fertility' was captured by Peirce with the phrase 'esperable uberty', applied to the 'hoped for' 'fruitfulness' or 'fertility' of scientific theories (see French, 1995). Peirce himself characterised this in terms of being '*gravid* with young truth' (1913).

The question then naturally arises, on what basis might we take such fertility in a theory to be 'hoped for'? And further, how might we evaluate whether a given theory is more or less fruitful than others? We suggest that the surprise evinced by certain consequences of the theory is one way of determining its 'esperable uberty'.

This seems evident in the case of $E = mc^2$, particularly given that the relationship between mass and energy had already been noted in the special case of the electromagnetic context. That it could be generalised through being derived from the theory of Special Relativity is indicative of the way that theory can be regarded as 'gravid with young truth'. And the hope that it would be fruitful was then confirmed by Meitner and Frisch's result.

Recent discussions of the value of such theoretical fertility have been shaped by McMullin's (1976) distinction between fertility in terms of the *actual* success that a theory has in opening up new avenues, dealing with problems and anomalies, etc., which he calls 'proven' or P-fertility; and fertility in the sense of designating the *potential* of a theory for future development, which he calls 'untested' or U-fertility. The former, of course, is retrospective, and is associated with the epistemic appraisal of a theory, being indicative of some degree of 'fit', again, between the theory and the relevant system (ibid., p. 400), whereas the latter is associated with its *heuristic* appraisal.

In the case of $E = mc^2$, we seem to have an obvious case of a move from 'U-fertility' in 1905, or 1909 in Meitner's case, to 'P-fertility' in 1939, the theory's fertility being 'proven' by the discovery of nuclear fission. Note, however, that it is not a case of overcoming some anomaly, as McMullin has it, but rather that of fertility manifested in terms of a 'new and powerful' extension of the theory of relativity. However, if U-fertility is taken to have only heuristic value then it and the surprise associated with the theoretical entailment prior to its confirmation and shift to P-fertility might appear to have no epistemic value at all (see Nolan, 1999). However, this ignores the 'esperable' or hoped for aspect. We recall, again, that although the extension of Einstein's theory was 'new' and hence might be regarded as the occasion for surprise, the relevant phenomenon, generally characterized, was not entirely

novel. As we have said, this supplied grounds for hope that the theory was indeed fertile. And this in turn suggests that, as with 'mere' surprise and confoundment we need to move beyond McMullin's classification, at least to some degree.

Consider: the 'potential' that a theory has for further development may be far ranging, covering all sorts of possibilities, from the trivial to the implausible. How should we determine which are indicative of the theory being 'gravid with young truth? Disruptive surprise may act as a 'flag' in such cases. Einstein's theory of Special Relativity was 'U-fertile' in all sorts of ways, of course; and its P-fertility was (eventually) demonstrable. But what McMullin's distinction fails to capture is the 'esperable' or hoped for fertility marked by the kinds of expressions of surprise we have noted here. Thus, this example nicely illustrates that the division between heuristic and epistemic appraisal may not be as clean as some might hope (see also da Costa & French, 2003, Ch. 6).[31]

7 Conclusion: The Disruptive Nature of Surprise

We began by considering Morgan's distinction between 'mere' surprise and confoundment, where the latter is distinguished from the former by virtue of the relevant result being inexplicable in terms of a given theory and thereby laying beyond our control. We have argued that, first of all, considerations of the value of surprise in science should be extended to thought experiments and theoretical derivations and secondly, that it is useful to see these as also more than 'merely' surprising and as disruptive, in a *productive* sense that is broader than confoundment.[32]

In all of the cases considered here we can tie the surprise involved to a certain disruptive feature. In the case of both Galileo's thought experiment and Einstein's theory of Special Relativity, the disruption was to pre-existing theoretical frameworks. It is perhaps almost trivial to describe the $E = mc^2$ result as disruptive, given the subsequent history. Nevertheless, it is worth noting that it can be seen as multiply so, beginning with remarks as to its 'cosmical importance' (Aston 1922), and continuing with the growing realisation of the implications of Frisch and Meitner's use of it. Just as scientists value the disruptive experimental results investigated by Ritson, so they value this aspect of both thought experiments, such as Galileo's and theoretical derivations, such as Einstein's.

[31] A referee has suggested that taking surprise as a mark of fruitfulness might be related to certain accounts of creativity (see, for example, Livingston, 2009; Thagard and Stewart, 2011). There may also be a connection to what Sheredos and Bechtel (2020) call 'imaginative success', whereby a possible mechanism is imagined that coheres with the available evidence and is taken to be hypothetically capable of producing a relevant explanandum relating to some phenomenon (it is then another step to determine whether that mechanism is actually responsible for that phenomenon). There is more to say here, particularly with regard to the role of the imagination, but we shall leave that for another occasion.

[32] Note that our arguments here do not require one to adopt a realist stance towards either theories or thought experiments. One might be an anti-realist of whatever kind and still maintain that surprise in general is valuable in science, not least as indicative of a certain fruitfulness, as we have indicated here, where that is disengaged from any notion of the theories that are developed as a result 'latching onto' the world, in whatever realist sense.

Of course, this is not the only respect in which surprise may have value although it may be the most pertinent in the theoretical context. And although it is difficult to conceive of phenomena in and of themselves as 'fertile' in this respect, one can surely extend the notion beyond the most theoretical levels to those typically described as 'phenomenological'.[33]

Our core claim, then, is that focussing on this disruptive aspect allows us to articulate an account of 'productive surprise' that accommodates surprising thought experiments and theoretical derivations. We suggest that this offers a broader and, as we have said, more useful perspective from which to view surprise in science.[34]

Acknowledgements We'd like to thank Aaron Meskin, Juha Saatsi, Adrian Currie, attendees at the Leeds HPS seminar and the Bristol Philosophy of Science seminar, and two anonymous referees for their helpful comments on this paper. Thank you also to Heinz Post for some initial inspiration.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Arcangeli, M. (2018). The hidden links between real, thought and numerical experiments. *Croatian Journal of Philosophy, 17*, 3–21.

Arthur, R. (1999). On thought experiments as a priori science. *International Studies in the Philosophy of Science, 13*, 215–229.

Aston, F. (1922). *Isotopes*. London: Arnold.

Ball, P. (2020). Behind the screens of AlphaFold. *Chemistry World* 9th December. https://www.chemistryworld.com/opinion/behind-the-screens-of-alphafold/4012867.article.

Bedessem, B., & Ruphy, S. (2019). Scientific autonomy and the unpredictability of scientific inquiry: The unexpected might not be where you would expect. *Studies in History and Philosophy of Science, 73*, 1–7.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science, 2*, 395–434.

Beisbart, C. (2018). Are computer simulations experiments? And if not, how are they related to each other? *European Journal for Philosophy of Science, 8*(2), 171–204.

Beisbart, C., & Norton, J. (2012). Why Monte Carlo simulations are inferences and not experiments. *International Studies in the Philosophy of Science, 26*(4), 403–422.

[33] Again, we are grateful to one of the referees for urging us to be more explicit here.

[34] This is not to say that 'matter doesn't matter' of course! The materiality of phenomena is significant, in various respects, but just not in the way Morgan takes it, when it comes to surprise, with her emphasis on the way the material world can confound. We recall Boyd's point that what matters is that scientific evidence should derive from a causal chain that is anchored in the world—what we are suggesting here, in effect, is that surprise conceived of as disruptive can arise at any point along this chain, and beyond.

Bokulich, A. (2001). Rethinking thought experiments. *Perspectives on Science, 9*(3), 285–307.
Boumans, M. (2012). Mathematics as quasi-matter to build models as instruments. In D. Dieks, W. J. Gonzalez, S. Hartmann, M. Stöltzner, & M. Weber (Eds.), *Probabilities, laws, and structures. The philosophy of science in a European perspective* (Vol. 3, pp. 307–318). Netherlands: Springer.
Boyd, N. (2018). Evidence enriched. *Philosophy of Science, 85*, 403–421.
Brendel, E. (2018). The argument view: are thought experiments mere picturesque arguments? In M. Stuart et al. (Eds.), *The routledge companion to thought experiments* (pp. 281–292). London: Routledge.
Brown, J. R. (1986). Thought experiments since the scientific revolution. *International Studies in the Philosophy of Science, 1*(1), 1–15.
Brown, J. R. (2007). Counter thought experiments. *Royal Institute of Philosophy Supplements*, *61*, 155–177.
Currie, A. (2018). The argument from surprise. *Canadian Journal of Philosophy, 48*(5), 639–661.
Da Costa, N., & French, S. (2003). *Science and partial truth: A unitary approach to models and scientific reasoning*. Oxford University Press.
Dardashti, R., Thébault, K., & Winsberg, E. (2017). Confirmation via analogue simulation: What dumb holes could tell us about gravity. *The British Journal for the Philosophy of Science, 68*, 55–89.
Di Paolo, E. A., Noble, J., & Bullock, S. (2000). Simulation models as opaque thought experiments. In M. A. Bedau, J. S. McCaskill, N. Packard, & S. Rasmussen (Eds.), *Seventh international conference on artificial life* (pp. 497–506). MIT Press.
Egeland, J. (2019). Imagination cannot justify empirical belief. *Episteme*. https://doi.org/10.1017/epi.2019.22.
Einstein, A. (1905). *Does the inertia of a body depend on its energy-content?* www.fourmilab.ch/etexts/einstein/E_mc2/e_mc2.pdf.
Feyerabend, P. (1975), *Against method*. New Left Books.
French, S. (1995). The esperable uberty of quantum chromodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 26*(1), 87–105.
French, S. (2020). *There are no such things as theories*. Oxford University Press.
French, S., and P. Vickers. (2011). Are there no such things as theories. *British Journal for the Philosophy of Science, 62*.
Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science, 49*(3), 397–424.
Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
Hitchcock, C., & Sober, E. (2004). Prediction verses accommodation and the risk of overfitting. *British Journal for the Philosophy of Science, 55*, 1–34.
Kind, A. (2018). How imagination gives rise to knowledge. In F. Macpherson & F. Dorsch (Eds.), *Perceptual imagination and perceptual memory*. Oxford University Press.
Lenhard, J. (2018). Thought experiments and simulation experiments: Exploring hypothetical worlds. In M. Stuart, et al. (Eds.), *The routledge companion to thought experiments* (pp. 484–497). London: Routledge.
Lenhard, J. (2019). *Calculated surprises: A philosophy of computer simulation*. Oxford University Press.
Livingston, P. (2009). Chapter seven. Poincaré's 'Delicate Sieve': On creativity and constraints in the arts. In K. Bardsley, D. Dutton, & M. Krausz (Eds.), *The idea of creativity* (pp. 127–146). Brill.
Mach, E. (1896). ber Gedankenexperimente. Translated by W. O. Price, S. Krimsky: On Thought Experiments, *Philosophical Forum 4*(3), 446–457 (1973).
McMullin, E. (1976). The fertility of theory and the unit for appraisal in science. *Boston studies in thephilosophy of science* (pp. 395–432).
Meitner, L. (1964). Looking back. *Bulletin of Atomic Scientists, 20*, 2–7.
Morgan, M. S. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology, 12*(2), 317–329.
Nolan, D. (1999). Is fertility virtuous in its own right? *The British Journal for the Philosophy of Science, 50*, 265–282.
Norton, J. D. (1991). Thought experiments in Einstein's work. In T. Horowitz & G.J. Massey (Eds.), *Thought experiments in science and philosophy*. Rowman & Littlefield.
Norton, J. D. (1996). Are thought experiments just what you thought? *Canadian Journal of Philosophy, 26*(3), 333–366.

Norton, S., & Suppe, F. (2001). Why atmospheric modeling is good science. In C. Miller & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 67–105). MIT Press.
Ohanian, H. C. (2009). Did Einstein prove E = mc^2? *Studies in History and Philosophy of Modern Physics, 40*, 167–173.
Parke, E. C. (2014). Experiments, simulations, and epistemic privilege. *Philosophy of Science, 81*(4), 516–536.
Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. *Synthese, 169*(3), 483–496.
Patricia Rife, P. (2019). *Lisa Meitner and the dawn of the nuclear age*. Plunkett Lake Press.
Peierls, R. (1979). *Surprises in theoretical physics*. Princeton University Press.
Peirce, C. S. (1913). An essay toward improving our reasoning in security and in uberty | EP 2:472. see http://www.commens.org/dictionary/term/uberty.
Poincaré, H. (1906). The end of matter (La Fin de la Matière, first published in *Athenæum* (1906), reprinted in "La Science et l'hypothèse" (edition from 1917, Chap. 14).
Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.
Popper, K. R. (1978). *Three worlds: The tanner lectures on human values*. Utah: Utah University Press.
Ritson, S. (2020). Probing novelty at the LHC: Heuristic appraisal of disruptive experimentation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 69*.
Sheredos, B., & Bechtel, W. (2020). Imagining mechanisms with diagrams. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination: Philosophical and psychological perspectives* (pp. 178–209). Oxford University Press.
Sime, R. L. (1997). *Lise Meitner a life in physics*. American Council of Learned Societies.
Sorensen, R. (1992). *Thought experiments*. Oxford University Press.
Stock, K. (2007). Sartre, Wittgenstein and learning from imagination. In P. Goldie & E. Schellekens (Eds.), *Philosophy and conceptual art* (pp. 171–194). Oxford University Press.
Stuart, M. T. (2016). Norton and the logic of thought experiments. *Axiomathes, 26*(4), 451–466.
Stuart, M. T. (2020). The productive anarchy of scientific imagination. *Philosophy of Science, 87*(5), 968–978.
Stuart, M. T., & Nersessian, N. (2019). Peeking inside the black box: A new kind of scientific visualization. *Minds and Machines, 29*, 87–107.
Thagard, P., & Stewart, T. (2011). The Aha! Experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1–33.
Todd, C. (2020). Imagination, aesthetic feelings, and scientific reasoning. In M. Ivanova & S. French (Eds.), *The aesthetics of science; beauty, imagination and understanding* (pp. 63–85). New York: Routledge.
Vickers, P. (2013). *Understanding inconsistent science*. Oxford University Press.
White, A. (1990). *The language of imagination*. Blackwell.