Synthese (2023) 201:34
https://doi.org/10.1007/s11229-022-04011-3

ORIGINAL RESEARCH

Putting the 'Experiment' back into the 'Thought Experiment'

Lorenzo Sartori[1]

Received: 10 June 2022 / Accepted: 14 December 2022

Abstract
Philosophers have debated at length the epistemological status of scientific thought experiments. I contend that the literature on this topic still lacks a common conceptual framework, a lacuna that produces radical disagreement among the participants in this debate. To remedy this problem, I suggest focusing on the distinction between the internal and the external validity of an experiment, which is also crucial for thought experiments. I then develop an account of both kinds of validity in the context of thought experiments. I show that we can naturally conceptualise internal validity in terms of games of make-believe. Then, I argue that external validity is best defined as accurate representation of a target system. Finally, I turn back to the current debate on thought experiments and show that my diarchic account provides a general framework that can be shared by the competing philosophical views, as well as a fruitful guide for their reconciliation.

Keywords Thought experiment · Internal validity · External validity · Experiment · Scientific representation · Make-believe

1 Introduction

In the last decades, philosophers of science have debated the epistemological status of scientific thought experiments (TEs):[1] how they function, what is their role in scientific investigation, and whether we can learn from them about the real world.

[1] If not specified otherwise, I restrict myself to scientific TEs. In my examples, I refer to TEs employed in physics, but work has been done on the role of TEs in other scientific disciplines, like economics (Schabas, 2018; Thoma, 2016) and biology (Schlaepfer & Weber, 2018), as well in non-entirely empirical disciplines like mathematics (Starikova & Giacquinto, 2018; see also examples of geometrical TEs in Brown, 2004) and philosophy. I hold that my argument works with all kinds of scientific TEs. Normative TEs definitely require further specification – cf. fn. 20.

✉ Lorenzo Sartori
L.Sartori1@lse.ac.uk

[1] Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, 28-30 Portugal St, London WC2A 2HE, UK

Published online: 19 January 2023

In Sect. 2, I reconstruct the major positions in the debate. In Sect. 3, I propose to reset the investigation on the basis of the similarity between thought experiments and material experiments (MEs). Specifically, I argue that we need to focus our attention on the distinction between the internal and the external validity of a thought experiment. Then, in Sect. 4 I offer a more detailed account of both these types of validity in thought experiments. I propose to analyse their internal validity in terms of Walton's games of make-believe, and to interpret external validity in terms of accurate representation. Finally, in Sect. 5 I go back to the current debate and show that the numerous positions presented in Sect. 2 can be explained and then reconciled with each other to a considerable extent.

2 An inflamed debate

2.1 Kuhn's questions

Scientific TEs[2] began to receive considerable attention from philosophers of science in the wake of Thomas Kuhn's (1977) provocative question: what is the epistemological status of TEs in science, given that they are apparently performed only in scientists' mind? Kuhn's question can be broken down in two sub-questions, which allow us to offer a first general taxonomy of philosophers' views on TEs:

1 Do TEs provide new knowledge about the empirical world?
2 If so, how do they do so? And if not, why not?

At this point, I do not need to commit to any particular theory of knowledge, and I take the term 'knowledge' to refer to whatever notion of knowledge is appropriate for science. The only constraint is that the purported knowledge concerns empirical facts, and when I say 'new', that the scientist acquires knowledge that they did not possess before performing the TE.

The first question allows us to divide positions in the debate up into two main camps, which I call the yes-camp and the no-camp. Positions in the first camp hold that we can achieve new knowledge via TEs and try to explain how this is possible. The positions in the opposite camp contend that TEs cannot provide new knowledge about the world, and give reasons to show why so. Within each camp, distinct positions can be further qualified by looking at how they answer the second question. This reconstruction of the many positions in the debate cannot be, and does not aim at being, exhaustive. Yet, framing the discussion on TEs from this particular perspective allows me to introduce and develop the issues of validity in the context of TEs.

2.2 The yes-camp and the no-camp

The yes-camp usually stems and acquires momentum from a historical perspective on TEs and their manifest role in modern science, from Galileo to Einstein, passing through Newton, Maxwell, and many other scientists. The yes-camp's leading intuition

[2] For a general overview, cf. (Brown and Fehige, 2022), (Stuart et al., 2018) and (Frappier et al., 2013).

is that TEs have importantly contributed to scientific progress. Further, as a springboard for the development of their views, most authors in this camp have emphasised, though in different ways, the peculiar role of the imagination in science and how it allows scientists to go beyond previous empirical knowledge. Among these positions, we can identify three main views, which I call for short Platonism, objectualism, and structuralism. Let us now see how they answer to the second question.

Brown (1992, 2004, 2011) argues that TEs can transcend the empirical knowledge that we possess before performing them. He contends that TEs contain an element of "a priori intuition" (2004, p. 31) that can neither be reduced to logical inference nor to the empirical knowledge already possessed by the scientist.[3] Particularly, TEs would allow us to "see" some of the general laws that govern our world. TEs are thus an instance of (potentially) powerful intuitive reasoning, which can make us immediately achieve some a priori content of knowledge about the world. Brown further characterises TEs' intuition in terms of a direct act of "perception" or "seeing", performed not through our senses by our mind itself, and directed to abstract objects, and specifically the very laws that govern nature (Brown, 2011, p. 98).

Platonism is not the only option to explain how TEs provide empirical knowledge. Gendler (2004) and Miščević (1992) , for example, focus on the objectual, picture-like nature of TEs imagination. This imagistic dimension of TEs, the authors contend, cannot be entirely reduced to a set of propositions; instead, by visualising an imaginary system, we acquire knowledge about how that very system would behave in the real world. This objectual nature of TEs' imagination is what makes them useful instruments of scientific investigation and explains how TEs provide new epistemic content. Nersessian (1992, 2007, 2018) has also proposed an overtly non-propositional view on TEs' imagination that, however, involves neither mental pictures nor visualisation. TEs would instead be instances of a *simulative model-based reasoning*, where a model is defined as a "structural analog" (1992, p. 293) of the scenario targeted by the TE. In Nersessian's terminology, a "structural analog" is a system isomorphic to the modelled system with respect to spatiotemporal and causal relations.

None of these accounts have been entirely successful in explaining the epistemic role of TEs in science. To begin with, the Platonic account appeals to somewhat mysterious processes of mental vision and a priori truths acquired via intuition. On such an account, it remains unclear how we can perform TEs that lead us to false results. Brown could say that, like sensorial perception, Platonic vision can go wrong too. However, while we have a scientific theory that explains how and why the former occurs, we have nothing comparable in the case of TEs. The objectualist view is intriguing, but it does not explain why the objectual nature of TEs' imagination makes an epistemic difference. TEs' perception-like nature may certainly make them effective in a pedagogical or heuristic sense, and their picture-like quality may offer different information, perhaps richer and more holistically integrated, than the one conveyed propositionally.[4] However, this proposal does not seem to provide any new insight on the acquisition of new empirical knowledge via TEs. For it is not real perception,

[3] The notion of intuition as a source of knowledge alternative to experience well fits the long-standing tradition of rationalism (see Markie, 2021, Sect. 2).

[4] Linguistic descriptions, mathematical formulae, diagrams, maps, photographs, graphs: all these types of representation have their own semantics and different ways to convey information. For example, pictures are

just imaginary one. And even if pictures and images have a different semantics than logically organised propositions, they are still created by our minds. So, either we picture things we already know, but then there is no epistemic gain through TEs; or objectual imagination allows us to achieve new content of knowledge, but this leads to exactly the same issues that we encounter when we elaborate new information from known propositions. So, we still have to specify how this acquisition of knowledge occurs, and on what basis the results obtained via the imagination are valid. All things considered, it would seem that there is not much of a difference between objectual and propositional imagination in this respect, and the burden to prove that an objectualist treatment is more successful than a propositional treatment lies on the objectualists' shoulders.

Nersessian's view seems to make some progress, but it requires an element of structural isomorphism between TEs and real systems that is both problematic and insufficient. Problematic, because it is difficult to understand why one should perform the TE in the first place, if one already knows the structure of the modelled system and the structure is all we need. Insufficient, because there are many different (and often irrelevant or misleading) ways a system can be structurally analogous to another one. The problem then is to find the *right* structure, among the many applicable to the same phenomenon. More generally, the appeal to non-propositional forms of imagination does not seem to answer Kuhn's second question satisfactorily, as from the point of view of knowledge (and not, say, feeling), any non-propositional content has a direct translation in propositional terms.

In contrast with the previous camp, philosophers in the no-camp have argued that TEs neither provide new knowledge about the empirical world, nor they actually can. Let us then introduce some views in this second camp.

On the extreme of the spectrum, we can place Dennett's view (1996) that TEs are simply "intuition pumps". Dennett seems to start from the same characterisation Brown gives of TEs, namely them being forms of intuitive reasoning, but he arrives at a radically different conclusion. He holds that TEs "are not arguments, they're stories". Instead of having a conclusion, "they pump an intuition" (p. 182). Exactly for this reason, they do not provide new knowledge about the world: they can just provide the initial intuition that needs to be supported by arguments and observations. A similarly minimalist view is held by Hacking when he says that TEs "can reveal tensions between one vision of the world and another" (1993, p. 307) and that they have no function beyond that. On this view, TEs are simply instruments to put our intuitions about reality into words or images. Hacking supports this conclusion by a comparison with material experiments (MEs). He contends that MEs acquire some form of autonomy from the experimenter: they offer new information about the world because they are themselves part of the world; they make it possible to perform new experiments; and they possess an element of surprise, as we do not have complete knowledge of the processes and mechanisms involved. TEs, instead, are inevitably

Footnote 4 continued
semantically dense in Goodman's (1976) terminology, and these make them particularly rich representations. See also (Camp, 2007) for maps, (Rescorla, 2009) for cognitive maps, (Perini, 2013) for diagrams, and (Perini, 2010) for a more general taxonomy of pictures' semiotics.

instances of the agent's imagination, and their relation with reality is mediated by the ideas and intuitions of the agent.

The answer that Dennett and Hacking give to the second question, i.e., why TEs cannot provide new knowledge, is not entirely satisfactory. Dennett just stops at the intuitive element and does not focus on the elaboration of information involved in thought experimentation. Eliciting intuitions does not seem the only thing we do when performing TEs. For we want to employ TEs to understand whether these intuitions are true or not, both in philosophy and in science. Hacking bases his argument on a comparison with MEs, but this is not going very far either: even admitting that TEs are not as independent of the experimenter as MEs, this does not entail that they cannot provide new knowledge about the world. Furthermore, (Shinod, 2017) has provided compelling arguments against Hacking's view, showing that TEs can exhibit a life on their own in Hacking's own terms, and they often produce surprising results.[5] In general, further unpacking is needed.

Norton (1991, 1996, 2004a, b), one of key contributors to the debate, provides a stronger answer to the second question. He starts from the empiricist premise that knowledge can be acquired only in two ways: by observation or by logical argument. As TEs do not involve the former, they must rely on the latter. Norton thus argues that TEs are logical arguments, presented in a picturesque fashion. Thus, TEs can also manifest interesting features of the real-world if they exhibit empirically grounded premises. However, they do not actually provide *new* knowledge about reality, as their empirical content was in fact already contained in the premises. TEs then turn out to be a rhetorically effective way to select and logically organise empirical premises in a salient way. Norton further defend his position by showing that one can effectively reconstruct successful TEs in the form of logical arguments with empirical premises. TEs epistemology turns out to be quite simple in Norton's eyes: a TE fails when it "is not sound", i.e., when the underlying argument exhibits at least one "false premiss or a fallacious inference" (2004b, p. 51).

Although Norton's account is compelling, it has some problematic implications,[6] the main one concerning the empirical premises required by his approach. For his account to work, we need TEs to be reconstructed as sound arguments about empirical facts. In order to be so, they need non-trivially true premises. However, it seems that many TEs involve unobserved, and sometimes unobservable things: someone running at the same speed as a beam of light, a lift away from any gravitational fields, minuscule demons, or, worse, compounds where a lighter stone acts as a brake on a heavier one. Of course, Norton can always offer an argument where these imaginative whimsies are swept away and still achieve the same results we obtained via the TE. But then one has the impression that we are not talking about TEs anymore, and the fact that they allowed scientists to formulate new ideas and gain knowledge starts looking like epistemic luck.

The last position I want to introduce has been defended by El Skaf (2018, 2021). Here, Hacking's characterisation of TEs as expression of a conflict between different visions of the world receives more qualification, with the simultaneous effort to avoid

[5] On TEs and surprise, see also (French & Murphy, 2021).

[6] For an exhaustive analysis of Norton's view on TEs, see (Brendel, 2018).

the most problematic issues of Norton's account. For El Skaf, TEs' primary purpose is to detect and sometimes resolve inconsistencies in our scientific theories. This can be achieved both intra-theoretically – when the TE is grounded in the tenets of a single theory – and inter-theoretically – that is, when the principles of two or more theories are combined. As an example of the former, El Skaf appeals to Galileo's TE on falling bodies.[7] An example of the latter is Bohr (1949) objection to Einstein's TE against the indeterminacy principle, where Bohr combines quantum mechanics with a relativistic framework.

For El Skaf, TEs do not in fact provide new epistemic content about the empirical world, as all the premises of the thought experimental scenario are derived from, or at least compatible with our scientific knowledge. TEs are then theoretical devices available to scientists to check, refine, amend, and sometimes abandon, their own scientific theories. Thus, either TEs' results remain *internal*, so to speak, to the theoretical domain, or they constitute possible models of those theories. In both cases, TEs concern theories, and they are concerned with the empirical world indirectly at best. So, we do not have to meet Norton's strong requirements of empirical truth: all we need is what our scientific theories tell us.

There are two problems with this view. First, it would seem controversial that *all* TEs are about theoretical inconsistencies. There is no inconsistency in Maxwell's demon, just a question about the nature of entropy. In Newton's thought experiment of two spheres rotating in an empty universe, there is no contradiction to be detected, but rather an abductive argument for the existence of absolute space. In Einstein's TE of a scientist trapped in a lift, the point is to show that uniformly accelerated motion is identical to the motion of bodies in free fall, not that their distinction leads to inconsistencies.[8] The second problem is the relation between the TEs and theories, on the one hand, and between TEs reality on the other. The purpose of TEs seems to tell us something about the empirical world, and in order to do that, they sometimes have to distance themselves from our established theories.

2.3 Internal debates and general problems

The positioning of authors with respect to Kuhn's questions has produced internal sub-debates between the protagonists of the discussion. First, Brown and Norton have discussed at length whether the knowledge achieved via TEs transcends empiricism – see, e.g., (Brown, 2004) and (Norton, 2004b). While Norton holds that TEs are just logical arguments in disguise, to be filled with available empirical knowledge, Brown argues that there is something more. Logic and empirical premises cannot account for the results we can achieve via thought experimentation, which forces us to postulate some form of a priori knowledge that is achieved via intuition.

Another discussion focused on the nature of the imagination in TEs, asking whether it is propositional or objectual – or, for Nersessian, at least not entirely propositional. This discussion revolves, again, around the epistemology of TEs. The arguments of the

[7] Galilei (1638), p. 107 of the original, p. 62 of the English translation.

[8] More on these TEs can be found in 3.2 and the references given there.

non-propositional view on imagination are motivated by the fact that propositions are not enough to explain the epistemic power of TEs, and their effective use in science.

Finally, as regards El Skaf's account, there is a question about the relation between TEs and scientific theories. He seems to hold that, if we learn something via TEs, it concerns what our theories imply. TEs' epistemic value would then strictly depend on theories, either because they inform relevant aspects of them, or because they show their shortcomings or problematic consequences. Naturally, this is also an attempt to answer the question concerning empiricism: there is actually no issue here, because TEs remain internal to the theoretical background. It seems also the most viable path for the no-camp, as it restricts TEs' knowledge to a purely theoretical one.

Now that the main positions have been introduced, and the sub-questions presented, I want to highlight two general problems. First, from the point of view of content, no account seems completely satisfactory. While the yes-camp remains quite obscure or vague in answering the second question, the no-camp risks to discard the epistemic value of TEs too quickly. Second, and more importantly, there seems to be a problem in the structure of the debate overall. As has become clear from our presentation of the state of the art, the problem with the philosophical debate on TEs is not only due to a plurality of epistemological accounts proposed. Rather, the different perspectives on TEs generate philosophical consequences in sharp contrast with each other. This should be clear from the sub-debates that I have just highlighted. All things considered, the impression neophytes inevitably have when they approach the debate is that there is little common ground shared by the discussants.

I contend that the reason for such a fundamental disagreement on TEs is that the community still lacks a theoretical framework shared by all participants. I suggest that a fruitful way to improve the current situation is to focus on the experimental nature of thought experimentation, and more specifically, on a fundamental but so far largely overlooked characteristic shared by both TEs and MEs: in both kinds of experiments, it is crucial to distinguish between the internal and the external validity of the experiment. This distinction is compatible in different ways with many views in the debate, and even anticipated in some works on TEs. Nevertheless, it has not been made explicit yet, or remained considerably underdeveloped. In the next section, after I have illustrated the general distinction for MEs, I show that it naturally applies to TEs. I conclude Sect. 3 by showing the novelty of this investigation with respect to other analyses of the experimental nature of TEs.

Once we put back the 'experiment' back into the 'thought experiment' by explicitly acknowledging the distinction between these two types of validity, and once we appreciate the deep consequences of this distinction, we acquire a more general conceptual background, so that we can better understand how the various positions in debate relate to each other. Consequently, the general epistemology of TEs becomes more limpid and more strongly related to the epistemology of experiments and scientific modelling. I will show this by both giving an answer to Kuhn's interrogatives and offering a way to solve the three sub-debates that I just have illustrated.

3 Two kinds of experimental validity

3.1 Validity in material experiments

There are two ways in which a material experiment can be valid: internally and externally. Campbell (1957) was the first to formulate such a distinction, and he associates internal validity with whether the experimenter's intervention is causally responsible for the observed outcome(s) occurring in the experimental system. By contrast, external validity concerns whether that very outcome can be generalised to other settings, and if so which ones.

More precisely, *internal* validity is the correctness of the experimental result in the original setting, which is studied by the experimenter under specified boundary conditions. Given an experimental setting *S*, constituted by a system *X* and a context (initial and boundary conditions) *C*:

> $Def_{(IV)}$: An internally valid experimental result *R* is a proposition made on the basis of the observation of, and/or an intervention on, *X*, which correctly ascribes a certain property *P* to *X* in *S*.

Thus, internal validity implies that the observed properties do not depend either on (i) the experimenters' mistakes or subjective biases, (ii) faulty measurement devices, (iii) a misinterpretation of the data (e.g., by taking a correlation as a direct sign of causation), or, most importantly, (iv) confounding factors, both expected factors and ones that had not been considered in the first place.

An experiment is *externally* valid when the internal results are successfully applied, or extrapolated, to other scenarios beyond the original experimental setting. Extrapolation here should be understood in a very general sense: induction from one specimen to other specimens; predictions about the outcome of an intervention when performed in a different context; hypothesising similar causal patterns in a new system; a statistical inference from a sample to a larger population.[9] Considering *S* as before, external validity can be defined as follows:

> $Def_{(EV)}$: An experimental result *R* is externally valid with respect to a designated target system *T* iff *R* is a proposition derived from the observation of, and/or an intervention on, an experimental system *X* in a setting *S*, and *R* is true about *T*.

Thus, the external validity of *R* is always relative to a specific target system *T*, which can differ from the observed and/or manipulated setting *S* on the basis of the system involved, the surrounding context, or both.

It is important to distinguish between these two types of validity because internal validity does not entail external validity. While the separability of the two is a general feature of experimental extrapolation, it has been emblematically acknowledged for so-called Randomised Controlled Trials (RCTs), a common methodology, particularly in medical research. RCTs are undoubtedly accepted as a gold standard for internal validity because their randomised sample selection usually succeeds in cancelling confounding causal factors out. For this very reason, however, it is difficult to assess

[9] Christensen and Waraczynski (1988, Sect. 4) propose a similar articulation of external validity.

when their results are successfully applicable outside the setting of the original trial, because we usually ignore what factors could be confounding or not. This results in serious issues for establishing external validity (cf. e.g., Cartwright, 2010b).

Consider the well-known case of the drug Benoxaprofen: although this drug had proven effective in several RCTs, it later turned out to be harmful when applied to the actual target group of patients (Worrall, 2007, pp. 994–995). The explanation of this mismatch is that while the tested groups included individuals of many different ages (indeed, a true random sample of the entire population), the actual target population was mostly composed by old people. Thus, the percentage of side effects, quite negligible in the random trials, was significant in the group of elderly people outside the laboratory walls. This is a clear example of a case where an internally valid result is not externally valid in the designated target.[10]

Another example of a mismatch between internal and external validity has been discussed by Cartwright (2012) and concerns the application of a plan for improving child nutrition. A policy focused on providing nutritional education to mothers had been introduced in a Northern region of India and was showing very promising results. Then, an attempt was made to export the same strategy to a similar problem of child nutrition in Bangladesh. This time, the intervention was a failure. The reason was that in Bangladesh, mothers had less control over food shopping than the mothers in Northern India, these decisions being more under the control of their mothers-in-law.[11]

Since Campbell's reflections, this distinction has generally framed the debate about the methodology of experiments.[12] The upshot of this excursion into MEs is that the there are two distinct types of validity of an experiment, and we have to keep this distinction in mind when it comes to understanding what knowledge an experiment can provide. Let me call this methodological thesis the Internal-External Validity Distinction (IEVD) and apply it to the current philosophical debate on scientific TEs.

3.2 Validity in thought experiments

I now give an example to show that it is useful to apply IEVD to TEs. Then, I will show that when the distinction is in place, it becomes clear that the best way to understand a TE's internal validity is in terms of games of make-believe, while the best way to interpret a TE's external validity is in terms of accurate representation.

In his *Dialogues Concerning Two New Sciences*, Galileo employs a TE to demonstrate the so-called law of equal heights.[13] Imagine a V-shaped cavity with a bottom

[10] There is nothing logically inconsistent in imagining the reverse case: an experiment mistakenly considered as internally valid that turned out to be externally valid. This is the reason for which, in $Def_{(EV)}$, I did not require internal validity as a necessary condition for the external one. However, it would certainly be unwise to build a methodology from this sort of cases, where external validity is acquired by mere luck.

[11] For other examples, see (Cartwright & Hardie, 2012).

[12] In experimental economics, cf. (Cartwright, 2007, Sect. 15) and (Guala, 2005, Sect. 7). Cartwright (1983) has also anticipated the same problems in physics; in psychology, see (Berkovitz & Donnerstein, 1982) and bibliography; in biochemistry, see (Strand et al., 1996).

[13] Cf. (Galilei, 1638), *Third Day*, Proposition XIII, p. 208. Particularly, see the comment to the figure at page 210. In the 1954 English translation, see pp. 216–218. This TE is also discussed by Salis and Frigg (2020, pp. 20–21) and Sorensen (1998, pp. 8–9).

that approximates a curve, to allow a ball to roll smoothly between the planes. Galileo demonstrates that if the ball rolls down the right plane of the cavity, it will reach the same height on the left plane, independently of the inclination of the two planes. For example, if one imagines bending the left plane downward, leaving the right side unchanged, the ball will still reach the same height on the left side. In modern physics, this follows directly from the conservation of energy. Galileo, instead, arrives at the law of equal heights via a long series of demonstrations, which concern the general features of uniformly accelerated motion. It is crucial to note that Galileo's derivation of the law strictly depends on some important idealisations. First, like most of the results obtained in Galileo's work on kinematics, this law depends on assuming the total absence of friction: the ball is perfectly spherical, the inclined planes are hard and smooth, and the air provides no resistance (Galileo, 1638, p. 166 in the original, p. 170 in the English translation). Second, specifically to this TE, once the ball reaches the vertex of the cavity and changes inclination of motion, it moves *as if* the conjunction of the two planes formed a curve. This second distortion remains implicit in Galileo's reasoning, but it is necessary because he restrains his analysis to accelerated motion along straight lines.

Immediately before presenting the TE, Galileo introduced what we now call *law of inertia*, which asserts that if no force acts on body, then it will either remain at rest or move with constant velocity. This was a significant change of paradigm in physics, as people previously thought that a moving object would grind to a halt once the cause of its motion ceased. The TE shows how the law of inertia follows from the law of equal heights. For, if one continues to bend the left side of the plane indefinitely, we obtain a case where it is actually horizontal. The law of equal heights states that the ball will not stop until it reaches the same height from which it fell. However, given that the plane is now completely flat, the ball never reaches the same height, and thus it must keep moving indefinitely. So, an object can move with uniform velocity with no force acting on it. The law of equal heights then is an intermediate step to demonstrate the law of inertia.

This perfectly fits our $Def_{(IV)}$: the law is an internally valid result as it correctly ascribes properties to the experimental system described by the thought experiment. However, the same result is not universally applicable, because in our world objects cannot move perpetually, and a ball rolling down one side of a cavity will not reach exactly the same height – only approximately so. Therefore, there is clearly a difference between a result that is valid within the thought experimental scenario, and a result which is also valid in a different context. Most importantly, the internal validity of the result, as it is evident in the case of Galileo's TE, does not *per se* entail external validity in any target T.

Galileo's case is no exception, and I contend that many other famous TEs are also fruitfully analysed through the lens of IEVD. For example, Maxwell considers a minuscule demon that is able to separate fast molecules from slow molecules in a box of gas.[14] The demon concentrates all of the fastest molecules in one half of the box, thus creating a considerable difference of temperature between the two sides of

[14] This TE was mentioned in a letter to Tait in 1867 (cf. Knott, 1911, pp. 213–215) and published in (Maxwell, 1871). Cf. (Norton, 2018) for an analysis of this TE and its recent reformulations.

the box. This scenario exemplifies a reduction of entropy in a closed physical system, and therefore counts as a counterexample to the second law of thermodynamics. This further means that the second law is statistical in nature: it is possible, just highly unlikely, that entropy decreases in a closed system. However, even though the result were valid in the TE, the applicability of this result to the external thermodynamic phenomena does not immediately follow – there are no such things as Maxwellian demons in the world.

Similarly, Newton asks us to imagine two spheres tied to each other with a rope in an otherwise empty universe.[15] The spheres are assumed to rotate around the common centre of mass, so each sphere is at rest relative to the other. If they are in motion, Newton argues, then there should be a force acting on the rope. Were the spheres at rest, this force would instead be absent. Newton suggests that this TE offers an argument against a relationist view of space. The relationist holds that only material objects exist, and space is just the set of spatial relations between them. Now, Newton argues, relationists are unable to offer any explanation for the presence of the force acting on the rope. This is because, from their point of view, motion is always relative to something else, and here the spheres are at rest with respect to each other. So, they do not move with respect to anything, and hence there can be no force between them. By contrast, the absolutist about space can provide an explanation for the presence of the force, namely that the two spheres are moving relative to absolute space itself and the force is not present when the two spheres are at rest, because they are not moving with respect to absolute space. Newton's TE has an abductive nature: absolute space would provide the best explanation for the presence of the force. However, this inference only concerns what is true within the TE. It is not evident per se that such a result is true about the nature of space in our actual universe, which is importantly different from the scenario described by Newton – for example, it is not empty.

Einstein's TE involves a scientist closed in a uniformly accelerated lift, which is not subject to any gravitational force.[16] In such a scenario, the scientist will see the objects moving as if they were subject to the effect of gravity. The TE's gist is that motion in accelerated frames of reference is identical to gravitational motion. With this, Einstein gets to posit the *principle of equivalence*, which establishes an identity between inertial and gravitational masses. This result applies to the imaginary, never observed scenario of a scientist trapped in a lift on which no gravitational force is acting. The point is then to show that what is true in the TE is also a thesis about real-world mass.

In scientific TEs, this passage from the imagined scenario to external targets is often mediated by theoretical and empirical assumptions. However, the inference from an imagined scenario to the behaviour of a real system is not as straightforward as it may seem at first glance. Cartwright highlighted this problem regarding what she calls Galilean experiments – MEs or TEs that isolate "a single factor as best as possible to observe its natural effect when it operates 'on its own' with no other causes at work" (2010a, p. 23). She focuses specifically on cases where an experiment involves unre-

[15] Isaac Newton, *Philosoph. Nat. princ. math.* (1687), *Definitiones*, 17, 11–12. Eng. transl. in (Cohen and Whitman 1999, pp. 414–415).

[16] Cf. (Einstein 2002, pp. 68–69) and (Einstein and Infeld, 1938, pp. 230–235). For an analysis, cf. (Norton, 1985).

alistic assumptions because in such cases it is necessary to "climb up the ladder of abstraction" in order to get "from falsehood to truth" (*ibid.*, p. 20). My attempt here is to provide a generalisation of Cartwright's point, namely a systematic account of the distinction between internal and external validity in thought experiments. Looking at this issue in relation to the case of MEs, one realises that the issues relative to extrapolation do not depend on unrealistic assumptions only. Sometimes, the assumptions made in the TE are realistic for some application, and unrealistic in other contexts. So, it is not simply a matter of falsehood and truth. Rather, I suggest it is better to interpret the issue in terms of internal and external validity. Specifically, I will argue in Sect. 4 that the passage from the experimental assumptions to the extrapolation to an external target is better understood as a matter of representation.

3.3 The experimental nature of TEs

This is not the first attempt to put the 'experiment' back into the 'thought experiment', in the general sense of relating TEs and MEs in order to reveal features of the former. Indeed, the literature on TEs contains numerous attempts of this sort. For this reason, I will briefly review the most relevant contributions on the experimental nature of TEs and show how my treatment either differs from, or goes beyond, previous proposals.[17] Readers unconcerned with matters of novelty can fast forward to Sect. 4.

Some of the contributions highlighting the relation between TEs and MEs insist on their mutual irreducibility or complementarity. For example, Sorensen (1998) draws important epistemological parallels between TEs and MEs, but his primary aim is to show how they serve different tasks, so concluding that the latter cannot entirely replace the former. In particular, he argues that TEs are examples of ideal experiments that are impossible to perform in the real world. In a similar vein, Buzzoni (2008, 2018) puts forward a transcendental interpretation of TEs, in which they would constitute the condition of possibility of MEs, both by framing the modal space of events and by providing a conceptual background to design and produce actual material experiments. Häggqvist also focuses on the modal features of TEs and claim that TEs are defined as hypothetical tests for theories. So, TEs do not directly provide knowledge about the empirical world (2009, pp. 59–60) – in this sense, he is close to El Skaf's account.

None of these three authors make explicit reference to validity. Furthermore, they all seem to share the core idea that TEs' results should be constrained when we turn our attention to the real world. If TEs do teach us something about the empirical world, it is just in the sense of delineating the possibility space for actual phenomena to occur.[18] I think that these accounts have problems in defining what "possible" exactly means, and to tailor it so that it can encompass all the rich varieties of "possibilities" drawn by TEs. In fact, many TEs employed in science describe scenarios that are just impossible. As Stuart (2020, pp. 972–976) points out, the imaginative activity in TEs, even scientific ones, is sometimes productive exactly because it is "anarchic", i.e., radically

[17] I thank an anonymous reviewer for pointing this out, as well as for suggesting most of the literature I address in this section.

[18] See also (Häggqvist, 2013) for how philosophical TEs do not by themselves provide justification for this sort of modal knowledge.

independent of previous theories and assumptions. This freedom can lead to imagine impossible scenarios, where physical laws are explicitly violated – think of Maxwell's demon, or even Einstein's scientist running at the same speed of a beam of light (cf. Norton, 2013). Moreover, these authors still express a unidimensional characterisation of TEs, without considering the two-pronged nature of validity emphasised in this paper.

Stuart (2016) proposes a "material" account of TEs, where the justification of their results depends not on the formal or logical relations between the propositions expressed by the TE, but on the material ones. He turns to Franklin's (1986) criteria for good material experiments in order to delineate analogous criteria for TEs. He takes into consideration the isolation of the experimental settings, the elimination of experimental bias, the identification of potential sources of error, the calibration of instruments, and the specification of a theory of measurement (Stuart, 2016, p. 460). While I agree with Stuart's application of these procedural guidelines to TEs, they seem to have little to do with IEVD. Indeed, some, if not all of these criteria have a different meaning depending on whether one is concerned with internal validity or with external validity. While isolating causal features can be relatively unconstrained in the internal dimension, the process of controlling causal factors will be more difficult when one applies the results to a real-world scenario. Different types of biases can affect the construction of the scenario and the inferences we draw about the external targets of our investigation. The theory of measurement Stuart proposes for TEs is a theory of inference making (*ibid.*, p. 461), but it disregards the fact that the inferences warranted within the TE may be very different from the ones concerning the world of phenomena, which is what we have to be more careful about. Hence, I take that Stuart's methodological recommendations remain orthogonal to my analysis.

IEVD should not be confused with other dichotomies. For example, the one between *interpretation* and *material realisation* suggested by Radder (1996, pp. 12–13) for MEs and applied by Mey (2003) to TEs. According to Radder, while interpretation concerns the outcome of an experiment *vis à vis* a precise theoretical background, material realisation is the idealised concept of an experiment *qua* a mere set of actions. Even if we grant appeal to a notion of an experimental action deprived of any theoretical interpretation, this distinction has nothing to do with IEVD. This is because both internal and external validity crucially depend on conceptual and theoretical assumptions.

Inspired by Mach (1896), Arcangeli (2018) distinguishes a dimension of production from a dimension of presentation in TEs. *The dimension of production* is the actual mental process of selecting and isolating features of the TE's scenario, manipulating of the imagined systems, and observing results. *The dimension of presentation* corresponds to the interpretation of the results in the light of a theory (*ibid.,* p. 17). Again, this distinction targets something very different from what IEVD targets. First, I take both the internal and external validity of a TE to depend on the scientific, theoretical background. Similarly to what I have said about Radder's distinction, it is not theories that get the lion share of the work in separating internal and external validity. Second, Arcangeli focuses on the dimension of production in order to show that the imagination involved in TEs is best characterised by appeal to mental models. This supports her view that TEs are useful because they allow us to "perceive" and "believe" from perspectives that are not directly present to our senses, and this constitutes the "experimental character" of TEs (*ibid.*, p. 15).[19] As I will show in Sect. 4.1, I offer a fictional treatment of the internal dimension of TEs, thus explicitly denying that belief is the cognitive attitude that the experimenter either does or indeed should entertain when they perform a TE. In this respect, Arcangeli's view is very different from my own account.

Let us now look at studies that have more or less explicitly appealed to IEVD. Wilson (2016), for example, has proposed such a distinction concerning moral TEs. However, his analysis does not give a precise account of either type of validity. More importantly, his analysis restricts itself to normative cases, which are relevantly different from factual cases as regards external validity. Even if there is not space to delve into this issue, the gist is that it is not clear whether we can think of the former in terms of representation at all – and if so, representation of what.[20]

In her doctoral dissertation, Murphy (2020) develops a rich comparison between TEs, MEs, and computer simulations. Although she is clearly aware of the fact that the IEVD can be drawn in all three activities, she does not give a precise account of both types of validity when it concerns TEs. She mentions external validity issues only when she criticises the alleged superiority of MEs with respect to computer simulations and, implicitly, TEs (cf. *ibid.*, pp. 33–37). The development I offer in the following sections, as well as the positive consequences of my treatment of TEs' validity in Sect. 5, thus go beyond Murphy's remarks while remaining compatible with them.

Finally, (El Skaf & Imbert, 2013) have offered a number of theses that are close to my own, though they differ in both their perspective and goal. They argue that TEs, MEs, and computer simulations share a "functional description", which is articulated as follows. They all are (i) question-oriented activities, and they involve (ii) a scenario, (iii) an unfolding of that scenario, (iv) the achievement of some results in the scenario, and finally (v) the obtaining of a scientific conclusion – i.e., an answer to the original question. While they neither talk about validity, nor refer to IEVD, El Skaf and Imbert clearly separate the results of the scenario from the answer to the scientific question. Thus, they have hit the same nerve on which I intend to focus, but without framing it in terms of validity.

The account proposed in this paper develops their ideas by being more general in certain respects, and by diverging from theirs in others. First, El Skaf and Imbert explicitly focus on the unfolding of the scenario, which corresponds to my internal dimension. My fictional treatment of internal validity is different and more general from theirs, insofar as I relate the internal dimension to the literature on fictions and imagination. Also, this fictional characterisation of TEs' scenarios distinguishes them in a relevant way from MEs and computer simulations, while it makes them more similar to scientific models. Furthermore, I offer a deep analysis of external validity, which is at best a secondary concern of theirs. I connect external validity with representation, which allows me to develop the view in a new direction. Their focus on the internal dimension leads them to say that the primary task of TEs is "explicatory" (*ibid.*, pp. 3463–3464) with respect to background theories, which means that their

[19] This bodily dimension, in that it allows us to feel, perceive, and thus believe, seems to be the main reason for which Arcangeli relates TEs to MEs also in other works, like her (2010, p. 584).

[20] Interesting thoughts on the same question, though applied to normative models, emerge in (Beck and Jahn, 2021) and (Roussos, 2020) .

primary function is to develop, analyse and assess theoretical assumptions. Instead, I want to insist on representation, and thus on the *external* relation between TEs and the world. Despite these differences, the fact that my account generally converges with El Skaf and Imbert's characterisation of TEs is a sign of how useful the distinction of validity can be at many different levels of the discussion.

Before I show how IEVD positively contributes to solving the controversy presented in Sect. 2, I need to spell out a precise account for both the internal and external validity of TEs, which I develop in the next section. This is a required step in order to fully appreciate the potential benefits of IEVD when applied to TEs, as well as to the issues currently troubling the relative philosophical debate.

4 Developing the account

4.1 Internal validity and games of make-believe

In this section, I propose an account for the internal validity of TEs in terms of Walton's (1990) games of make-believe. The benefits of connecting TEs to Walton's treatment of artworks have been noted before. In particular, Meynell (2014) is the first to explicitly suggest interpreting TEs in this way. However, as it often happens with fictionalist approaches to scientific contexts, Meynell does not offer a precise account of how exactly TEs provide knowledge about the empirical world (cf. *ibid.*, p. 4165), thereby failing to address the relevant problem of TEs, as identified by Kuhn. Instead, Meynell seems satisfied with Walton's account as providing a definitive answer to the question regarding the epistemological status of TEs – or, at least, providing the fundamental ground for any philosophical analysis of them. I intend to argue that Walton's game semantics do indeed provide a neat account of the internal validity of a TE. However, in contrast with Meynell, I contend that this approach gives us just half the story; further work is needed to analyse the external validity of TEs, and to address the related issue of how TEs may offer new knowledge about the empirical world.

Salis and Frigg (2020) also employ Walton's games for an analysis of the scientific imagination involved in TEs and scientific models. The view put forward by these authors is more similar to mine than to Meynell's, insofar as they restrict themselves to the internal dimension of imagination, without developing their account any further about the possible epistemic import of TEs for the empirical world. In this section, I first introduce the broad outlines of a Waltonian account of TEs on the basis of the previous works I have just mentioned, and then apply it as a general framework for the internal validity of TEs.

The concept of game of make-believe has been originally introduced by Walton in order to explain how artistic representation works. When we look at a piece of art, we are engaging in a game, where the material elements of the artwork have to be interpreted following specific rules. For example, most bi-dimensional coloured canvases must be interpreted as presenting three-dimensional objects; a blindfolded woman with a scale in her hand should be understood as an allegory of justice, and so on. This also applies to non-pictorial arts, like literature. Abbott's *Flatland*, for example, tells the story of a two-dimensional object that encounters objects from a three-dimensional reality. Here, the written text and the pictures in the book prompt our imagination to create a fictional scenario and develop the narrative. In Walton's terminology, the material vehicle of the game, like pictures and written texts, is called the *prop*, while the rules that guide the construction of the game's scenario and its further unfolding are called *principles of generation*.

The game activity thus generates a fictional scenario, or fictional world, which can be defined on the basis of the propositions that are true in it (Walton, 1990, pp. 35 ff.).[21] Let us call 'w-fictional' a proposition that is true in a game of make-believe w. A proposition then is w-fictional iff it is directly expressed by the prop (the writing, or the image) or derived from it through the principles of generation assumed in w. Salis and Frigg (2020, p. 35) call the former set of truths *primary truths*, and the latter *implied* or *derivative truths*. Derivative truths may not be explicitly stated in the description of the fictional scenario. They are inferred by further reasoning, on the basis of prop together with the principles of generation that are in play in that specific game.

This approach, as Frigg and Salis have already shown, neatly captures the key features of the scientific imagination, i.e., the kind of imagination involved in scientific models and TEs. In order to show how this works, let us return to Galileo's TE from the previous section. One has a text, written by Galileo himself, that works as a prop for our imaginative activity. Further, this activity is governed by explicit but also implicit assumptions and rules of inference, given by the scientific context in which we are operating. The primary truths provided by the prop concern the existence of one ball on the edge of a V-shaped cavity and the idealised features of these objects, as well as the absence of friction. The explicit principles of generation at play are: the definition of uniformly accelerated motion; the absence of friction; and further idealisations about the motion of the ball when it approaches the vertex. However, other principles are in place, for example classic mathematical derivations and logical inferences. As it is evident at this point, we are prescribed to imagine a scenario where false statements and true ones are irremediably intertwined. From the combination of prop and principles we obtain the derivative truth that, once the ball starts rolling down the slope, it will reach the same height on the other side of the cavity. Furthermore, it is true in the fictional scenario that, were one of the arms of the cavity bent till being horizontal, the ball would proceed to roll forever, thus exemplifying the modern law of inertia.

This view can just as easily be applied to most TEs used in science: Maxwell's demon in the box, the imaginary scientist that Einstein conceived trapped in an elevator, Newton's two lonely spheres rotating in an otherwise empty universe, and so on. In all these cases, a prop and a set of principles of generation can be identified. It is possible to further infer the derivative truths to understand the properties exemplified by the TE's scenario. Accordingly, we can say that a result is internally valid in the thought experimental setting w if and only if it is w-fictional, that is, it is part of the explicit description of the fictional world, or it is derived via a principle of generation.

Sometimes, internal validity is not easy to establish. For example, Mach (1919, pp. 228–238) contests that there is no way, in Newton's TE, to ascertain that the rope

[21] For reasons of space, I put the question of the ontology of fictions aside: my notion of fiction is compatible with different ontological interpretations, including anti-realist ones.

is actually undergoing that force. For there is no observable difference between the case where the spheres are at rest and the case where the spheres are rotating. In other words, Mach is accusing Newton of begging the question: he is already assuming the negation of the relationist thesis, namely, that there is any physical difference between the two cases. In my framework, this concerns the internal validity of the TE, as it concerns what is true in the thought experimental scenario. Here, I do not wish to take a stance on who is right or wrong in the controversy. What I want to show is only that this should be conceived of as a debate on the internal validity of Newton's TE, and that one can analyse the internal tenability of a TE by investigating what is actually true in the imaginary system.

In a Waltonian framework, there can be countless derivative truths, which all have the potential to be relevant results for subsequent scientific investigations. In this sense, the account is different from the one proposed by El Skaf and Imbert (2013): they insist on a prominent difference between the unfolding of the (TE's) scenario and the internal results. In contrast, my account has no principled way to identify "the" results. All derivative truths are on the same level and can in principle play the role of internal results. What counts as a "result" of the TE will depend on the context of external application, once we try to extrapolate the properties of the surrogative system to real targets.[22]

It is important to insist on the intrinsically normative nature of Waltonian games (Walton, 2020, p. 36). The game prescribes us to imagine certain contents and rules others out. We cannot ignore the fact that the protagonist of *Flatland* is a square. Similarly, when we entertain Galileo's TE game of make-believe, we have to assume the definition of uniformly accelerated motion. Inversely, we are not allowed to imagine that the square in *Flatland* has five sides, or that there is friction between objects in Galileo's TE. More generally, there are licit and illicit acts of imagination. These are determined more or less explicitly by the principles of imagination, which can be dependent on the constraints an epistemic community holds regarding the specific context of investigation. Thus, the results of our interpretation of, and reasoning about, the fictional scenario can be evaluated on the basis of the legitimacy of the individual imagining. This process becomes even more rigorous in the scientific imagination, where the principles of generation are inferential schemes, mathematical theorems, evidence-based assumptions, and tenets of our most general theories. In other words, the fact that we are imagining does not mean that anything goes.

It is also crucial to highlight that what is true in a fictional world is independent of the concept of truth *simpliciter*, whatever theory of truth one may want to endorse. What is true or false in the game is solely determined by the prop and the assumed principles of generation. The fact that there has actually never been a fabulous treasure on the little island of Monte Cristo, for example, is just irrelevant for the game we play when we read Dumas' book *The Count of Monte Cristo*. The same holds for cases of scientific imagination, like Galileo's TE. In fact, the scientific imagination often presents a mixture of truth and falsehood, with observation-based elements merging

[22] I think that El Skaf and Imbert will have to refer to external factors in order to distinguish the results from the general unfolding of the scenario. For example, they may appeal to the scientific question the TE is meant to answer. I want to resist this, because the internal dimension should retain enough independence of the specifics of the external application.

with theoretical assumptions, idealisations, and abstractions. Fiction should therefore not be confused neither with truth nor falsity (cf. also Frigg & Nguyen, 2020, Ch. 6). In addition, this semantic independence of Walton's games with respect to truth has an important consequence at the epistemic level, namely that one is neither committed to believing in any particular content of the game, nor to believe that such content is false. So, what is required to be imagined has no bearing on the credential attitude we ought to take towards it, and the prescriptive character of Waltonian imagination does not entail any kind of epistemic commitment.

All things considered, there are two main advantages of treating the internal validity of TEs in this way. First, this approach allows enough freedom to employ false assumptions without requiring that our attitude to them is belief. Consequently, the account also keeps explicitly distinct the internal level of analysis from the external one. As we are not concerned with truth *simpliciter* from the start, we refrain ourselves from making any inference about external phenomena. Second, once the game is on, it imposes strict rules, which simultaneously captures the normative dimension of the scientific imagination, its social dimension, and its potential for rigour. Thus, the rules of the game endow it with a prescriptive nature, balanced against the freedom of imagination.

The make-believe account, when applied to the scientific imagination, has been the target of many criticisms. Friend (2020) and Thomasson (2020) have cast doubts on the ability of fictions to denote and to be elements of comparisons with real world systems. Todd (2020) has also raised some doubts on Salis and Frigg's proposal: if all we achieve from TEs and scientific models already depends on the principles we started with, how can we learn something new? The make-believe account also has to compete against alternative accounts – e.g., Godfrey-Smith (2020) treats the scientific imagination in terms of counterfactual conditionals, and French (2020) characterises TEs credence status in terms of "quasi-truth".

In defence of the make-believe account, it must first be noted that much of the general criticism raised against the fiction view of scientific models is not relevant to the present work, because my aim is to restrict Walton's account to what concerns internal validity only. Therefore, I grant that this view is in itself limited and demands to be integrated with an account of the epistemic relation with the external world. Such an account will be offered in Sects. 4.2 and 4.3. Besides, even if the objections raised against the fiction view were relevant to this restricted application, much work has already been done to answer them. For example, Frigg and Nguyen (2021) debunk several commonplaces about the fiction view on models. Salis et al. (2020) also offer important reflections on how scientific fictions can denote, without renouncing to the basic anti-realist tenets of their account of fiction. Moreover, Salis (2016) offers a way to make sense of fiction-world comparisons in a fictionalist framework. Furthermore, in their paper, Salis and Frigg (2020) do not only introduce the make-believe account but also give compelling reasons for why it is preferable to other treatments (like the counterfactual one), and why it is important to clearly distinguish the imaginative attitude from the belief attitude.[23] Concerning Todd's remarks, I think that they may

[23] This last point is in continuity with the remarks offered by Stuart (2020): we must be free to explore different points of views and apparently counterintuitive ideas in order to progress in our scientific

be misplaced: scientists are not logically omniscient, so even if they knew all the initial assumptions that govern the imaginary system, this would not rule out the possibility of them being genuinely surprised by the results they achieve by studying the scenario's implications.[24] At the same time, scientists may not be able to explicitly list from the start all the necessary and sufficient assumptions required for our scenario to work. It is also the task of the philosopher of science to analyse scientific fictions and reconstruct the relevant assumptions at play.

Finally, my aim here is not to defend Walton's view *per se*, but to show that it provides an optimal way to address the problem of TEs' internal validity. Thus, my thesis here should be taken as purely conditional: if one applies Walton's make-believe semantics, then one achieves a working account of internal validity for scientific TEs. There is unfortunately no space here to delve into the debate on fictionalism any further. However, at least abductively speaking, the fact that Walton's account provides a neat answer to the definition of internal validity in scientific thought experiments should already count as a good argument for considering the view seriously.

Because of the intrinsic epistemic restrictions that I have been applying to the fictionalist approach, I have yet to clarify the relation between the scientific imagination and the knowledge about the empirical world. In my framework, this concerns the problem of external validity of TEs and is the topic of the next two sections.

4.2 TEs as representations

TEs employed in science are, of course, not only games of make-believe.[25] *Qua* instruments of scientific investigation, they can also be evaluated on the basis of their efficacy in providing us with information about the external world.

We often want to use TEs as a means of surrogative reasoning about the world. Galileo's TE with the V-shaped plane is not just a speculative exercise to show that the law of inertia is true in *that* fictional scenario: it is meant to be an argument for the truth of that law in the actual world. Then, to paraphrase (Brown, 2011), the game of make-believe of Sect. 4.1 becomes an optimal "laboratory of ideas", where theoretical hypotheses, empirical observation and purely fictional elements interact fruitfully, the ultimate goal being the discovery of interesting features of reality.

A question then arises about the validity of reasoning about an external target when performed on the basis of a TE. Ultimately, this is the central question for the epistemology of TEs: whether our imaginative activity can lead us to knowledge about the world. I propose to preliminarily distinguish two aspects of the question that have often been merged in the debate on the epistemology of TEs. One question concerns the *definition* of external validity; another, the *methods* one can employ to assess external

Footnote 23 continued
understanding, and this makes much more sense if our imaginative activities do not require us to epistemically commit to the fictions we entertain in our minds.

[24] It is not always just a matter of lack of logical omniscience: cf. (French & Murphy, 2021) and the element of genuine surprise of TEs.

[25] Nor often are artworks: paintings and novels are often telling about the real world. For example, *Flatland* is meant to be a mordant parody of the hypocrisy and closed-mindedness of the Victorian society, and Orwell's *Animal Farm* should be read as an allegory of Stalin's regime.

validity or alternatively, the justification of claims as externally valid.[26] An answer to the definitional question, I argue below, does not determinate a universal answer to the methodological question. However, the way in which we define external validity will have important consequences for our way to assess it.

Let us start with the definitional question. Instead of imposing the definition of external validity from the outset, I suggest looking at Galileo's TE once again and see how it relates to an external target. This TE describes a fictional scenario that, once interpreted as a surrogative system, is meant to highlight some specific properties possessed by real physical systems. The TE therefore refers to, or *denotes*, some real target systems in the world. Moreover, the TE *exemplifies* the relevant properties to be attributed to the target. Exemplification here is meant in the technical sense as intended by Elgin (1983, 1996) and Goodman (1976) and then by Frigg and Nguyen (2020, pp. 172–174). An object X exemplifies the property *A* iff (i) *X* possesses *A* and (ii) *X* refers to *A* in a context *C*. The notion of exemplification also involves making features salient and epistemically accessible (Elgin, 1996, Ch. 6). Consequently, some properties are more salient than others, which could also be abstracted away or distorted. Galileo's TE exemplifies many interesting properties, among which the law of equal heights and the law of inertia. It distorts the vertex of the cavity and ignores friction.

Now, these properties are intended to be *imputed* to systems in the world. For example, one may want to say that the V-shaped cavity is to represent the oscillatory motion of a pendulum, or the motion of rolling objects on a curved surface – like a skateboarder on a two-sided ramp. The law of inertia is meant to be a general feature of motion in our world, expressed in the form of a counterfactual statement. In this sense, Galileo's TE establishes the following property: if an object were not subject to any force, it would persist in its state of motion (i.e., it would either be at rest or move in a straight line with constant speed). This counterfactual[27] statement is true about all moving systems in our world. However, it does not need to be counterfactual, as objects in interstellar space come pretty close to being described by this law too. In this sense, the law is re-adapted so it can be applied to real physical systems via approximation, and not just counterfactually. So, the properties exemplified by a TE map onto different targets in different ways.

Frigg and Nguyen (2020, pp. 174–176) call this process of translating the properties of the representation into the properties of the target system *keying-up*. The key is thought of as a function, mapping the properties exemplified by the representation onto the properties that we actually want to impute to the target. You can think of keys as the one we find with maps, specifying how to read them correctly. In Galileo's case, we have a key that translates a factual statement into a counterfactual one, but also a key that works as an approximation.

The four terms italicised correspond to the four crucial elements defining the relation between the TE and its target: denotation, exemplification, keying-up, and imputation.

[26] The same holds for internal validity: one will define it in terms of fictional truths, and then assess whether a claim is internally valid by checking whether it follows from the prop combined with the principles of generation.

[27] A counterfactual interpretation seems the most natural way to go. At the same time, this seems close to Nguyen's (2020) appeal to "susceptibility" when he describes the application of results that we achieve from extremely idealised models.

These are also the fundamental ingredients of the so-called DEKI account of scientific representation, developed by Frigg and Nguyen (2020, pp. 159–215). DEKI is in fact an acronym, standing for denotation, exemplification, keying-up and imputation. Given the analysis of Galileo's TE just offered, it is natural to interpret this TE as a representation of mechanic motion in the real world. This is exactly the suggestion I want to put forward, namely, to interpret TEs as representations in the sense expressed by DEKI, and to define external validity as accurate representation of a target system.

DEKI's concept of representation is useful because it allows us to conceive of TEs that misrepresent the target. Sometimes a TE's result is simply false when applied to the intended target.[28] An account of representation based on a relation of similarity – or its formalised version, isomorphism – would instead not permit this solution. This is because, whatever similarity is, either two things are similar, or they are not.[29] Furthermore, DEKI allows thought experimental scenarios *in their entirety* to be very different from any real target. Of course, the target scenario can in principle be identical to the imagined one. Then, the TE's results are trivially true about the target as well. However, these cases are rare, because the very point of a TE is to investigate aspects of reality that we do not know yet. In fact, we can seldom be aware, in advance and with certainty, of all the relevant features of the target, particularly the ones we want to reveal via a TE. Accordingly, DEKI underlines the importance of both exemplification and of keying-up: exemplification, because TEs can highlight some properties at the expense of others that are overshadowed or distorted; keying-up, because the properties exemplified by a surrogative system usually need a translating function that maps them onto the intended target. We have already seen this with the different ways the law of inertia is mapped onto different classes of target systems. So, there is an evident need for different keys, depending on the specifics of the designated target. This does not only mean that the imputed properties may vary depending on the target: they are also patently different from the ones literally exemplified by the TE.[30]

Besides what we do not yet know, DEKI also sheds light on the fact that TEs can involve elements that *we know* are just false when applied to their target. If what is needed in the first place is a true description of the target in order to assess its external validity, then Galileo failed from the outset, regardless of the kind of extrapolation employed. For example, objects are not perfectly smooth in the real world. This is again accommodated by the functioning of exemplification: in order to emphasise some properties, others will inevitably be omitted from consideration. This aspect is also in accordance with the freedom implied by a fictional treatment of TEs internal validity.[31]

[28] The history of science is full of examples of this sort. Lucretius' (*De rerum natura* I, 968–983) argument against the idea of a finite universe is one of them.

[29] In addition, as Goodman (1976) famously noticed, similarity and isomorphism are symmetric, reflexive, and transitive, whereas representation is not (cf. Suárez, 2003). See also (Frigg, 2006).

[30] If one were troubled by the idea of fictional scenarios instantiating properties, there are good news: the model system instantiates properties only as an object interpreted by some function I. So, the properties are I-instantiated and therefore I-exemplified. See (Frigg & Nguyen, 2016, p. 228).

[31] In fact, Frigg and Nguyen themselves endorse a Waltonian view of models as fictions, and they use it to integrate their account by drawing parallels with Goodman's idea of Z-representation (see Frigg & Nguyen, 2016).

My analysis seems to naturally account for the functioning of Galileo's TE. It also fits the mould of all the other TEs that I mentioned so far. For example, Newton's TE aims at establishing a property of real space, namely its independence of the objects inhabiting it, by exemplifying that property. Here, the key keeps the property of being absolute unchanged. In the case of Maxwell's TE, the actual contradiction of the second law of thermodynamics is best to be converted into a statistical interpretation of the same law. Finally, concerning Einstein's lift, nobody has in fact ever observed a uniformly accelerated box away from any gravitational field. The observational identity between uniformly accelerated motion and the motion in a gravitational field has to be translated into a true identity between the two types of motions. This gap between what is true in the TE and what is true in the world follows from the very characteristics of TEs as fictional scenarios employed to represent external targets. While there are prescriptions about the content to be imagined, nothing yet forces us to believe the contents of our imaginings as true. At the same time, TEs are still intended as tools to understand something about the empirical world. I contend that they do it by exemplifying certain properties, which are then applied to a target system in the real world via a proper key.

Note that all the aspects that I have illustrated so far regarding TEs also apply to scientific models. Models are often employed to discover features of targets about which we are still unaware or uncertain; they usually include plainly false assumptions or highly idealised controls; and these assumptions are not just a necessary evil, they also positively contribute to the success of the model's external validity. Moreover, despite not commonly being framed in terms of internal and external validity, the literature on scientific models offers a great deal of analysis on the distinction between truth within the model and truth about the target.[32] Given these relevant similarities, I suggest addressing the issue of TEs' external validity in close analogy with how we address related questions in the context of scientific modelling.[33] At this point, the reader may suspect that I am just identifying TEs and models. However, I contend that such an equation is not warranted on the basis of what I have said so far. I take TEs and models to share a distinction between internal and external validity, but this holds for MEs as well, when employed for extrapolation. This does not imply that MEs, TEs and models are all the same thing. After all, Walton's account functions well with both scientific surrogative systems and with works of art, but this does not entail that art fictions are exactly the same as scientific one, even within the internal dimension: the principles of generation will diverge to a considerable extent. Finally, the DEKI account of representation applies to many different types of representations: maps, diagrams, scans, simulations, material and theoretical models. Nevertheless, the generality of DEKI should then not be understood as implying an identity between all these different types of surrogative systems.

[32] This distinction is a basic tenet of the so-called representation-as accounts (Elgin 1983, 1996 and Goodman, 1976) and of the DEKI account (Frigg & Nguyen, 2020, Sect. 8). Similar intuitions are expressed also in (Hughes, 1997, p. S332) and (Tan, 2021, pp. 16–18).

[33] Salis and Frigg (2020) have shown that the internal activity of imagination conducted in TEs and in scientific models is fundamentally the same. Here, I am completing the picture by showing the similarity between TEs and models when it comes to external validity.

In general, I put the question about the relation between models and TEs aside. I simply intend to look at how one deals with external validity in models and take inspiration for what concerns TEs. Specifically, I want to focus on the idea that models facilitate successful surrogative reasoning about their targets by means of *accurately* representing them.[34] I suggest that the same holds for TEs. This is the topic of the next section, where I propose to define external validity of TEs in terms of accurate representation.

4.3 External validity as accurate representation

I take it that the accuracy of a representation is always relative to two factors: the designated target of the representation, and the specific set of properties that is eventually imputed to that target. So, we can never talk of an accurate representation *simpliciter*: we have to specify the target and what property exemplified by the representation we want to impute to the target – once properly translated via a key. Let me first propose a more precise definition of what accurate representation is, building on the DEKI account of representation:

> *Accurate representation*: A is an accurate representation of a designated target T regarding a set of properties Q iff (1) A exemplifies a set of properties P; (2) P is converted via a proper key into Q;[35] (3) Q is imputed to T; and (4) T actually possesses Q.

On the basis of this, we can give a clear definition of external validity for TEs:

> *TE external validity*: a TE is externally valid with respect to a designated target T relative to a set of properties Q iff the TE is an accurate representation of T relative to Q.

Thus, I suggest that the extent to which our TE-based extrapolations are valid depends on whether the TE exemplifies properties that, once translated via the appropriate key, are correctly ascribed to the designated target. We need a key in order to address the fact that, depending on the target, the same property can map in numerous ways from the same TE to distinct target systems. As I highlighted above, the law of inertia applies in different ways depending on the type of target – sometimes counterfactually, sometimes as an approximation. Of course, the key should not be completely *ad hoc*, but rather it should associate the properties of the TE and the target in a systematic way. This is the case in Galileo's TE and the other examples mentioned so far.

It is important to stress here that my definition of accurate representation, which offers the ground for my definition of external validity of TEs, does not involve any assumption about the *similarity* between the representation system and its target. Indeed, accuracy here only depends on the application, via a proper key, of *some*

[34] I acknowledge that there are views which undermine the role of representation in models functioning (see, e.g., Isaac, 2013). A structured reply to this lies beyond the scope of this article. However, I have on my side (i) common scientific practice – scientists investigate phenomena indirectly via models on an everyday basis; and (ii) broad philosophical work – among many, see (Frigg & Nguyen, 2020) and (Weisberg 2013).

[35] The key may also be a relation of identity, mapping P onto itself.

results obtained from the study of the representation. The key included in my definition, inherited from the DEKI account, allows for very different, and sometimes purely conventional ways to connect the properties of the representation with the properties of the target.[36]

The fact that accuracy is independent of similarity makes my account compatible with, if not in fact a theoretical ground for, what (Stuart, 2020) calls the "productive anarchy" of TEs.[37] Stuart argues that TEs are sometimes useful exactly because they challenge our theories and intuitions in a revolutionary, radical way. Consequently, they often involve scenarios that are extremely different from the usual ones, if not even impossible according to our best scientific theories. Despite their anarchic nature, these TEs are still important, Stuart argues, because they make us critically reflect on our theories and unchallenged intuitions. My definition of accurate representation gives us another good reason for not being too troubled about TEs' "anarchy": a representation can be accurate, despite its lack of realism or even of a similarity with its targets. Therefore, we do not need to renounce accuracy in order to account for the revolutionary nature of some TEs.

Let us now discuss possible types of targets of TEs. The target of a TE can of course be one single object. For example, I take Newton's TE with the two rotating spheres to target physical space. Yet, this is rarely the case, as TEs normally tend to represent classes of systems or types of mechanisms – what Weisberg (2013, Sect. 7) calls "non-specific targets". As stated before, Galileo's TE represents pendulum-like motions, and the law of inertia is counterfactually targeting any motion in the world. Similarly, Einstein's lift is a representation of a type of motion, specifically the one instantiated by objects subject to gravitational fields and by objects moving with uniform acceleration. Maxwell's TE targets closed thermodynamic systems and exemplifies the statistical nature of a vast set of phenomena occurring at the level of the fundamental components of matter. This should not worry us, as TEs are no exception in this respect: scientific models usually target broad, and sometimes even vaguely defined, classes of phenomena. Bohr's model of the atom, for example, is not supposed to be the representation of a particular atom but of the class of all hydrogen atoms.

Furthermore, nothing in my definition rules out the possibility of a targetless TE whose goal is a purely theoretical investigation. For instance, one may wonder if the famous TE that Galileo offers in his *Dialogues* (cf. *infra*, fn. 4) against Aristotle's theory of falling bodies is actually a representation of anything. The TE is meant to be a *reductio ad absurdum* of Aristotle and his followers' theory of falling bodies, which states that heavier bodies fall faster than lighter ones. But what if, Galileo reasons, we connect one heavy object L and a lighter object l with a rope, and we let them fall from Pisa's tower? If the fictional system satisfies the laws of the Aristotelian doctrine of motion, then we obtain two mutually contradictory answers. On the one hand, the

[36] By detaching accurate representation from similarity, I am in total agreement with Nguyen (2020, Sect. 4) and with what he calls "interpretational" accounts of scientific representation, among which he also lists DEKI. For a survey of these accounts, see (Frigg, 2022, Ch. 9).

[37] Similar ideas are expressed in (Murphy, 2022, Sect. 4). I thank an anonymous reviewer for pointing me to a possible tension between my insistence on accuracy and the emphasis that some authors put on the freedom or anarchy involved in TEs.

lighter object l should act as a brake, so that the resulting velocity of the compound is somewhere in between l's and L's velocity. On the other hand, l+L is itself one single compound, which is heavier than L alone. So, the compound's velocity must be strictly greater than L's one. Therefore, either Aristotle's theory is contradictory, or it is vague enough to produce contradictions.[38]

This TE may be said to be targetless: it is only a fictional scenario employed to reflect on the implications of Aristotle's theory.[39] Alternatively, one may argue that Galileo's falling bodies has a target after all, however general and vague it may be. In fact, we would be imputing the property that the speed of falling objects does not depend on their weight. Brown (2004, pp. 30–31) goes even further and contends that there is a general claim we can make about reality on the basis of this TE. Namely, that it is impossible that the velocity of a falling body depends on an extensive property of the object – that is, properties that can be added and subtracted as if they were real numbers.

I am not taking position on this specific question, as it concerns the independent problem of tackling representations with vague targets.[40] What matters is that, even in cases of targetless TEs, the framework that I have sketched still applies. In fact, one can still talk about representation here: adopting (Goodman, 1976) terminology, they are Z-representations. Even though a painting of a unicorn is targetless, as there are no unicorns, we can still consider the painting as a representation, namely a unicorn-representation. In the same way, Galileo's TE on falling bodies is an Aristotelian-falling-bodies-representation, with possibly no targets in the world.

4.4 The justification of external validity

Once external validity of a TE is defined, we need to discuss the issue of how a scientists can assess whether a TE is externally valid. This is not a question about what external validity is, but rather *how* we find out whether a TE is accurately representing something or not. In other words: what is the epistemology of external validity and how do we justify our inferences from the surrogative system to the target one?

There is no ready-made recipe to determine whether a TE is externally valid. In this TEs are like other surrogative system such as models or MEs aiming at extrapolation. Think of an experiment performed, say, on mice in order to study the neurological mechanisms of memory. The experimental results are externally valid with respect to, e.g., humans iff what we find out in mice turns out to be true in humans as well. How do we know that the inference is justified? The answer largely lies outside of the experiment itself, and the methods to provide it can be numerous: performing further experiments on different organisms; establishing parallelisms between mice brains and human ones; evaluation, via archaeological and genetic data, of the tenability of phylogenetic assumptions about rodents and primates. Furthermore, all these forms of

[38] For a thorough analysis of this TE, see (El Skaf, 2018) and (Gendler, 1998).

[39] Thus, this TE would work more like a model of a theory than a model of phenomena, because it provides a scenario that simply tries to make Aristotle's tenets true. See (Frigg & Hartmann, 2020) for the distinction.

[40] As Frigg and Nguyen (2020, pp. 13–14) notice, the very question of whether a model is targetless or not is tricky, as it strongly depends on how the target is defined.

investigation may find foundation in overarching theories – in this case, the contemporary theory of evolution – the justification of which again relies upon observations, experiments, and other theories.

Of course, even if we performed many experiments, severe tests of our hypotheses, and robust analyses of our measurements, this will not necessarily entail that we will have achieved a definitive justification of the extrapolation. In principle, the process of justification can go on indefinitely. There will simply be a moment when the scientific community will reach a provisional consensus on the tenability of the required assumptions. The same holds for models: there is no one-size-fits-all method to decide whether a model's results are externally valid about a target system, just by looking at the model itself. Fine-grained analysis must be carried out on the theoretical assumptions involved in the model, together with empirical investigation. This is the reason why some models' external validity is so difficult to assess. In some cases, it is because past empirical data are not entirely sufficient for justification – e.g., consider the case of climate models, where the effects that humanity started to have on the climate around 200 years ago make the data about farther past less relevant. In other cases, it is the overarching theories that are not a matter of consensus yet – e.g., in economics and psychology – so they are not sufficient to justify the model-based inferences. Generally, the grounds to justify both experimental extrapolations and model-based inferences partially lie outside them. The surrogative system of course participates in the justification of the extrapolation, but it is in itself insufficient to ground it completely.

The partially extrinsic nature of justification holds for scientific TEs as well. Take once again Galileo's cavity. We are warranted in applying the result of the TE to real motion because some assumptions are empirically grounded, the idealisations are known to function well in that context of application, and other empirical evidence (e.g., on pendulums) indicate the same conclusions. All this is just extrinsic to the TE itself. Again, this does not undermine our definition of external validity in terms of accurate representation. It just shows that sometimes it is difficult to assess whether a representation is accurate or not.

At this point, one may complain that TEs seem much more distant from empirical reality than scientific models or MEs. Thus, we may be not allowed to answer the problem of justification simply by associating TEs with the other two types of surrogative reasoning.[41] Irrespective of how distance is interpreted here, this will be a matter of degree. If we understand distance in terms of similarity, we can have both very "realistic" TEs and very idealised ones, as well as more or less artificial experiments; analogously, we can have both very unrealistic models, and models that describe accurately many aspects of the target. So, there does not seem to be an essential difference between TEs, models and MEs in terms of their distance from external targets. Whatever surrogative system one employs, an extrapolation to a target will always require justification. In addition, as I have already emphasised, the definition of accuracy that I am suggesting is independent of the similarity with the target. Therefore, we should not worry about the lack of realism of most TEs, because they can be accurate representations of their targets even when they appear unrealistic.

[41] I thank an anonymous reviewer for pointing this out.

One final caveat concerns the relation between the internal and the external dimension. It is important to clarify that IEVD does not imply that the imaginary scenario and its representational function have nothing to do with each other. When it comes to actually constructing TEs, scientists will obviously make considerations of empirical and theoretical nature. Therefore, they will aim from the start at accurately representing something in the world. This does not undermine IEVD, because the two types of validity, although being intertwined in practice, remain conceptually distinct.

In conclusion, I do not think that a definitive set of universal criteria to assess the accuracy of a representation can be established. Like in the case of MEs and scientific models, the success of exporting information from the surrogative system to the target one can be assessed only a posteriori. As in all cases of reasoning based on surrogative systems, results are always tentative and conjectural, just as any other scientific result. In this sense again, we can appreciate the true experimental nature of TEs.

5 Stabilising the debate

5.1 Amending the yes-no debate

We can now answer Kuhn's questions: TEs can produce knowledge about the empirical world by providing true propositions about real target systems. These true propositions are justified by both the TE itself, and theoretical assumptions with their empirical support. This latter ground of justification remains importantly extrinsic to the TE itself. However, I have argued that this is not problematic in principle, as the same occurs in the case of model-based inferences and experimental extrapolation. From this perspective, the problem of justifying the external validity of TEs is just a special case of the broad problem of justifying scientific inferences, and this can be solved only by (i) looking at the background scientific knowledge more holistically, and (ii) taking into consideration the specific context in which the inference is performed.

In addition to giving an answer to Kuhn's questions, IEVD and the consequent diarchic account of TEs' validity also allow us to re-frame the debate on the epistemological status of TEs. Let us start with the yes-camp, and specifically with Brown. He has tried to answer Kuhn's questions by focusing on the internal dimension of TEs. His account tends to lump together all the aspects that I have distinguished in this article: internal validity, external validity, and their justification collapse and remain within the TE's scenario itself. Through imaginative activity, Brown argues, we can infer not only what is true in the TE's scenario, but also about the real world, namely true laws of nature. The warrant of our inference is given in both cases by the strength and immediacy of our intuition. If one accepts my distinction between internal validity and external validity, this is too quick. For we cannot automatically infer external validity from the internal one. One has to recognise that there is a fundamental difference between what occurs in the scenario, and how the target system behaves in the world.

One may want to redefine Brown's account in the light of my distinction. We first find out what is true in the TE, under the assumption that some laws govern the scenario; and *if* the same laws also govern our world, then the TE is externally valid. Although Brown does not seem to have this in mind, as he argues that the laws themselves are discovered via the TE (2004, p. 34), I find this reinterpretation of Brown's account more appealing. At the same time, I find any reference to laws of nature, a priori intuition, and Platonic perception unnecessary. My account thus seems preferable, insofar as it does not require laws to be the same in the surrogative system and in the target system, even though this is certainly the case in many scientific TEs. At the same time, Brown's main intuition is retained: we can learn something about the real world via thought experimentation, and this occurs by interpreting TEs as representations in DEKI's sense.

With respect to the yes-camp in general, my account provides a more qualified answer to how we obtain knowledge about the world. The flexibility of the DEKI account is useful to capture the different ways in which a TE can relate to the external target. Compare, for example, the representational account of external validity with the objectualist accounts. As I noticed above, it is not entirely clear why the picture-like nature of TEs imagination should add something crucial for our TEs to succeed. Instead, my representational view accounts for it and allows for both propositional and non-propositional treatments of imagination. This is because the account holds that the TE, *qua* representation, exemplifies certain properties, and it includes the use of keys to properly translate TEs' properties into features of real targets. The way in which the selection of exemplified features is achieved will of course depend on the type of representation: pictures and images have syntactic and semantic features that may be important to understand how some properties are exemplified in certain TEs. Here, the work of Goodman (1976) and Perini (2010) helps highlighting the peculiarity of visual and in general non-propositional representations. For example, picture-like images tend to be syntactically and semantically dense, thus making them rich in detail while remaining concise. Also, spatial relations play an important role as representing other forms of relations. All these features may be relevant to understand how some TEs exemplify properties that are then imputed to a target. My point is that, first, when we are concerned with validity, a distinction between visual and propositional representation does not help us much at the general level, but rather only at the local one of exemplification. Second, at least when one takes into consideration examples of TEs in physics, they do not seem to require an appeal to specifically non-propositional features to be valid, either internally or externally.

The advantages of the account are evident also with respect to Nersessian's model-based account. In a sense, Nersessian's idea that a TE is a structural analog of a real system is similar to the idea of external validity that I put forward here. However, her concept of structural analog does not provide a satisfactory account of scientific TEs' validity. For the isomorphism is not sufficient for a TE to be externally valid. In fact, many things we imagine are often isomorphic in some way to external systems, but this does not make them externally valid. This is because isomorphism is a very abstract notion that can be easily instantiated. We need a richer conceptual background to tell us which isomorphism is the relevant one. In this sense, Nersessian's structural analogy is not a sufficient condition for external validity. Furthermore, being a structural analogy does not seem to be particularly useful to understand the internal validity of TEs, and thus investigate their imaginary dimension. For one may want to perform TEs that are not externally valid but only internally so, like Galileo's TE on falling bodies. Here, it is at least unclear to what the fictional system is isomorphic, given that nothing in nature seems to instantiate the contradiction exemplified by Galileo's fictional system.

More generally, it seems that a structural analogy pertains to a possible method of justifying the external validity of TEs. In fact, it can be part of the explanation of why a TE is an accurate representation of a target. This method of justification of external validity, though, should neither be confused with external validity itself, nor with the TE *tout court*. My account can incorporate Nersessian's view as a strategy to justify TE-based inferences. Then, isomorphisms will usually have to be further qualified by an interpretation of the fictional system, importantly extrinsic to the fictional scenario itself, which will explain *in what sense* it is relevantly isomorphic to real ones. In other words, we will need a key, and thus some further work is required to motivate the choice of the designated isomorphism in stead of other ones.

As regards the no-camp, my account firstly explains why TEs can be intuition pumps in the first place. This is because they offer a free space for our imagination to combine true elements with fictional ones, opening up in front of us possibilities and hidden aspects of the investigation that may have remained implicit or unknown. It also adds crucial information about why this is relevant for science, by providing an analysis of the relation between TEs and external world in terms of representation.

About this last point, El Skaf thinks that we usually do not get "outside" TEs' internal domain. Thus, the only epistemic function TEs can serve is a theoretical one. More specifically, they reveal inconsistencies in our theories (and potential solutions to them). My reply to this view is twofold. First, even remaining within the internal domain, we do not need to restrict ourselves to inconsistencies. Galileo's cavity does not bring up any contradiction, just the implications of some general assumptions combined with empirical data and some relevant idealisations. Similarly, Newton's spheres, Einstein's lift, and Maxwell's demon are not meant to identify contradictions in our scientific theories, but rather to make their consequences manifest. Imagination allows us to stretch the limits of our theoretical knowledge, not only to challenge it. Furthermore, even though TEs instantiate many theoretical assumptions, they are not entirely reducible to them, given the presence of fictional particulars and idealisations. Second, while the results of TEs may sometimes be only internal (e.g., Galileo's falling bodies), this is not necessarily the case for all scientific TEs. Actually, most of them are meant to give us information about the real world. If we think TEs in terms of DEKI, we can account for this fact in terms of surrogative reasoning. Therefore, El Skaf's view can also be incorporated in my general framework, as it describes the special case of TEs that reveal and solve contradictions in our theories.

Finally, let us turn to Norton. His method consists in getting rid of the imaginary elements and reconstructing the TE as a logical argument, where the premises express empirical knowledge or well-grounded theoretical claims. Therefore, trying to explain TEs' external validity, he imposes constraints on internal validity. Furthermore, by identifying TEs with underlying logical arguments based on empirical premises, he also seems to conflate external validity with the method employed to justify it. As a consequence, he has to reject the view that there is something new that we discover via TEs because all the empirical content is already present in the premises already.

I contend that this strategy is problematic. To begin with, his account is unable to explain for the importance of imagination and fiction, which play a crucial role in

TEs. In fact, fictional elements are vital to achieve the internal results, and also to make salient those properties that we want to impute to the target. Without assuming imaginary objects and their behaviour, the TE normally does not work. In Galileo's cavity, as almost always in physics, we deal with idealised objects, the behaviour of which is rarely if ever approximated in the empirical world. Salis and Frigg (2020, p. 37) have shown that the same holds in Galileo's TE on falling bodies: without assuming events that actually never obtain in the world, Galileo would be unable to prove the inconsistency of Aristotle's theory. The same holds for Einstein's scientist running parallel to a beam of light (Stuart, 2020).

One may wonder whether this is actually in contrast with Norton, who sometimes allows some role for imagination and the particular elements described in the narrative. Fiction would then be useful because it facilitates reasoning. I have two main reasons for doubting that Norton's view in its current formulation is able to take this route. First, while Norton understands fictions and imaginary particulars as merely allowed, I intend to stress that TEs are usually epistemically valid, in the internal sense, exactly *because* they involve idealisations, abstractions, approximations, false assumptions, and so on. Moreover, even when it comes to external application, idealisations are not just a necessary evil: like models, TEs use these distortions fruitfully in order to make some properties salient at the expense of others. Without idealised assumptions on the absence of friction, Galileo would have been incapable of formulating the law of inertia. Therefore, we actually want fictional particulars and idealisations: they are not synonyms of inaccuracy,[42] they are an essential element of scientific enquiry.

Second, the main argument that Norton gives for the superiority of his account is that it gives a clear method to choose between TEs with mutually contradictory results. He calls these "thought experiment – anti thought experiment pairs" (2004b, p. 45). If two TEs have mutually contradictory results, we need a systematic way to understand which one is the correct one. Now, Norton's account envisages two possibilities: either (at least) one of the two underlying arguments is logically invalid, or (at least) one has false premises. However, the first option is quite rare. So, his solution to the problem, which is put forward as a crucial reason to prefer his account, basically relies on the assumption that a good TE has true premises. However, if Norton allows for fictions and imaginary elements to play a role in the scenario, then he loses this option, and his account is not better off than those that he criticises.

My view offers a neat solution to this problem: Norton just needs to split his account in two and consider each TE as (usually) composed by *two* arguments: one is internal to the scenario and ruled by the chosen principles of generations, the other concerns the extrapolation of results from the scenario to an external target. In this way, we can allow scientists to use fictions in the internal dimension, while our empiricist tenets can be retained in the argument we give for external validity. The challenge then becomes to see whether there is a good key to translate the exemplified properties into ones to be imputed successfully.

Finally, it is worth noticing that my proposal is in line with other authors' concerns about Norton's account. For example, Stuart (2016) argues that Norton seems forced to

[42] This very point is made by Nguyen (2020) concerning toy models.

renounce at least one of following: (i) TEs provide knowledge, (ii) an empiricist theory of TEs justification, or (iii) Norton's material theory of induction (Norton, 2021):

> There is therefore a serious internal tension between Norton's account of thought experiments according to which thought experiments are filled with irrelevant but picturesque details on the one hand, and his account of induction according to which the particular details are crucially important for justification, on the other [...] it would be instructive to consider whether another empiricist or naturalist account of thought experiments can be created that explains their ultimate source of justification (Stuart 2016, pp. 458–459).

Stuart then goes on in specifying the criteria for a good TE that I summarised in 3.3. I contend that my proposal provides the empiricist account of TEs that Stuart hopes for. Following my distinction of internal and external validity, I acknowledge the importance of imaginary particulars for what concerns the internal results, while at the same time I place the justification of the external applications on theoretical assumptions and empirical knowledge, which remain largely extrinsic to the TE itself *qua* representation. In this sense, also the material theory of induction seems to be retained, given the justificatory role of background, material knowledge, both empirical and theoretical, that has already been taken for granted by scientists. Finally, my proposal permits us to see how TEs provide new knowledge: they can produce new justified true beliefs about their targets by accurately representing them.

What I have said does not undermine in any way the importance of Norton's work. By offering a rational reconstruction of TEs as an argument, Norton is normally able to assess their external (and sometimes internal) validity. What I have argued is simply that this method of assessing TEs' external validity should not be confused with TEs themselves, and we should not collapse internal and external validity. In the end, Norton's argument view is perfectly adaptable to my IEVD: it is sufficient to acknowledge that there are usually two arguments usually involved in TEs. Moreover, I think that my account adds something, as it gives a more qualified characterisation of what *type* of "arguments" are involved in the two cases, respectively: one is a game of pretence, the other a representation-wise inference. In this sense, I am providing the material aspects to complete Norton's formalist-empiricist account.

5.2 Remedying the sub-debates

Now, I can also show how IEVD and my diarchic account help mitigating the contrast developed by the protagonists of the debate. First of all, with my distinction in place, we can both retain Norton's empiricism and accommodate Brown's concerns. Given the freedom, autonomy, and even the anarchy allowed in a game of make-believe, a TE can include elements that go well beyond background theories, play with possible and impossible situations, and mix them creatively with both empirical observations and intuitions. The point is that the internally valid results are not unconditionally valid for any external target: whether they are valid or not depends on their accuracy, which in turn depends on the specification of the exemplified properties, the designated target, and the key involved. Moreover, the justification of the key is importantly extrinsic to the single TE and rely on theoretical and empirical knowledge, and this is in perfect harmony with Norton's empiricist requirements.

Furthermore, the discussion about propositional vs. objectual imagination finds a more precise place into the debate. Once we recognise that TEs work as representations in DEKI's sense, both propositional and objectual imagination can be accounted. The difference between picture-like and linguistic representations, in terms of their syntactic and semantic peculiarities (cf. aforementioned works by Goodman and Perini), will become important to understand a specific "step" of the representation process, namely exemplification. However, despite the importance of this aspect, it is clear now that a focus on the non-propositional nature of some representations does not solve the questions of whether and how TEs generally produce knowledge about the empirical world. My account succeeds in that by highlighting more general features of TE-based reasoning. In addition, at least in the case of the TEs in physics that I take as case studies, there seems to be no reason to appeal to non-linguistic representational features. A reference to these peculiarly non-propositional properties could nevertheless be useful for other examples of TEs, or different kinds of reasoning.

Finally, in line with what I said about TEs and empiricism, we have a valid alternative to El Skaf's strategy to restrict TEs' results to the internal, theoretical domain. For IEVD gives us a way to distinguish between two different interpretations of TEs' results and allow the externally valid results to be different from the internally valid results. In addition, the appeal to the concept of representation in DEKI's terms allows us to re-connect TEs with the empirical world in a straightforward way. A further, interesting consequence of my account is that TEs can thus sometimes be relevantly independent of our scientific theories. Like scientific models, they will then be able to play an autonomous, auxiliary role in bridging the gap between theories and phenomena. In this way, we are able to answer El Skaf's concerns optimistically.

6 Conclusion

In this article, I proposed to re-frame the debate on TEs on the basis of the distinction between internal and external validity, borrowed from the parallel distinction employed in the philosophical literature about material experiments. I illustrate the distinction by analysing Galileo's thought experiment of a ball rolling in a V-shaped cavity. Then, I provide two detailed accounts of internal and external validity of TEs, respectively. I suggest that we should think of the former in terms of Walton's games of make-believe, and that the valid exportation of the internal results of TEs to external-world contexts is best interpreted as a process of accurate representation. On the basis of this diarchic account, I have provided an answer to Kuhn's initial questions: TEs are games of make-believe that provide knowledge about the real world by representing their targets accurately. Fictions are best interpreted in terms of Walton's games of make-believe, and the concept of accurate representation is clarified by the use of the DEKI account of representation. Finally, my account offers the opportunity to re-interpret previous positions and disagreement by providing a common conceptual framework. With my diarchic account of TEs' validity in place, I can both do justice to the different voices introduced in Sect. 2 and establish a fruitful dialogue between them. Especially, the account explains the reasons supporting both camps, while at the same time it gives an escape route to the impasses produced by the radically different views animating the debate.

Acknowledgements This work would have never seen the light of day without the ceaseless help of Roman Frigg, to whom I express all my gratitude. I also want to thank two anonymous reviewers and Rawad El Skaf for all their invaluable comments. Finally, many thanks to Michael Stuart, James Nguyen, Joe Roussos, Michal Hladky, Lukas Beck, Marcel Jahn, and all the members of the Working Models Reading Group.

Declarations

Declarations I confirm that I have no conflict of interest and that I adhere to all the requirements of the COPE guidance.

References

Arcangeli, M. (2010). Imagination in thought experimentation: Sketching a cognitive approach to thought experiments. In L. Magnani, W. Carnielli, & C. Pizzi (Eds.), *Model-based reasoning in science and technology* (pp. 571–587). Springer.

Arcangeli, M. (2018). The hidden links between real, thought and numerical experiments. *Croatian Journal of Philosophy, 18*(1), 3–22.

Beck, L., & Jahn, M. (2021). Normative models and their success. *Philosophy of the Social Sciences, 51*(2), 123–150.

Berkovitz, L., & Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist, 37*(3), 245–57.

Bohr, N. (1949). Discussion with Einstein on epistemological problems in atomic physics. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-scientist* (Vol. 42, pp. 199–242). The Library of Living Philosophers.

Brendel, E. (2018). The argument view: Are thought experiments mere picturesque arguments? In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 23–43). Routledge.

Brown, J. R. (1992) Why empiricism won't work. In P. A. Schilpp (Eds.), *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Number 2 in 1992, (pp. 271–279). Philosophy of Science Association.

Brown, J. R. (2004). Why thought experiments transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 23–43). Blackwell.

Brown, J. R. (2011). *The laboratory of the mind: Thought experiments in the natural sciences*. Routledge.

Brown, J. R., & Fehige, Y. (2022). Thought experiments. The Stanford Encyclopedia of Philosophy.

Buzzoni, M. (2008). *Thought experiment in the natural sciences*. Königshausen and Neumann: An operational and reflective-transcendental conception.

Buzzoni, M. (2018). Kantian accounts of thought experiments. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 327–341). Routledge.

Camp, E. (2007). Thinking with maps. Philosophy of. *Mind, 21*, 145–182.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Cartwright, N. (2010). Models: Parables v fables. In R. Frigg & M. Hunter (Eds.), *Beyond mimesis and convention: Representation in art and science* (pp. 19–31). Springer.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies, 147*(1), 59–70.

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), 973–989.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Christensen, L. B., & Waraczynski, M. A. (1988). *Experimental methodology*. Allyn & Bacon.

Cohen, I. B., & Whitman, A. (1999). Isaac Newton. *Principia*: Mathematical principles of natural philosophy. University of California Press.

Dennett, D. (1996). Intuition pumps. In J. Brockman (Ed.), *Third culture: Beyond the scientific revolution* (pp. 181–197). Simon and Schuster.

Einstein, A. (2002). *Relativity: The Special and the General Theory (1916)* (Robert W. Lawson Eng. Trans.). Routledge.

Einstein, A., & Infeld, L. (1938). *The evolution of physics*. Cambridge University Press.

El Skaf, R. (2018). The function and limit of Galileo's falling bodies thought experiment: Absolute weight, specific weight and the medium's resistance. *Croatian Journal of Philosophy, 18*(52), 37–58.

El Skaf, R. (2021). Probing theoretical statements with thought experiments. *Synthese, 199*(3), 1–29.

El Skaf, R., & Imbert, C. (2013). Unfolding in the empirical sciences: Experiments, thought experiments and computer simulations. *Synthese, 190*(16), 3451–3474.

Elgin, C. Z. (1983). *With reference to reference*. Hackett.

Elgin, C. Z. (1996). *Considered judgement*. Princeton University Press.

Franklin, A. (1986). *The neglect of experiment*. Cambridge University Press.

Frappier, M., Meynell, L., & Brown, J. R. (2013). *Thought experiments in philosophy, science, and the arts*. Routledge.

French, S. (2020). Imagination in scientific practice. *European Journal for Philosophy of Science, 10*(27), 1–19.

French, S., & Murphy, A. (2021). The value of surprise in science. *Erkenntnis, 2021*, 1–20. https://doi.org/10.1007/s10670-021-00410-z

Friend, S. (2020). The fictional character of scientific models. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination* (pp. 122–127). Oxford University Press.

Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria, 21*(55), 49–65.

Frigg, R. (2022). *Models and theories*. Routledge.

Frigg, R., & Hartmann, S. (2020). Models in science. The Stanford Encyclopedia of Philosophy.

Frigg, R., & Nguyen, J. (2016). The fiction view of models reloaded. *The Monist, 99*(3), 251–269.

Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*. Springer.

Frigg, R., & Nguyen, J. (2021). Seven myths about the fiction view of models. In A. Cassini & J. Redmond (Eds.), *Models and idealizations in science* (pp. 133–157). Springer.

Galilei, G. (1638). *Discorsi e dimostrazioni matematiche intorno a due nuove scienze* (H. Crew & A. de Salvio Eng. Trans.). New York: Dover Publications. (Original work published 1954).

Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science, 49*(3), 397–424.

Gendler, T. S. (2004). Thought experiments rethought – and reperceived. *Philosophy of Science, 71*(5), 1154–1163.

Godfrey-Smith, P. (2020). Models, fictions, and conditionals. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination* (pp. 154–177). Oxford University Press.

Goodman, N. (1976). *Languages of art*. Hackett.

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

Hacking, I. (1993). Do thought experiments have a life of their own? Comments on James Brown, Nancy Nersessian and David Gooding. In Hull, D., M. Forbes, & K. Okruhlik (Eds.), *Proceedings of the Philosophy of Science Association Conference 1992*, Volume 2 (pp. 291–301). Chicago: University of Chicago Press.

Häggqvist, S. (2009). A model for thought experiments. *Canadian Journal of Philosophy, 39*(1), 55–76.

Häggqvist, S. (2013). Modal knowledge and the form of thought experiments. In A. Casullo & J. C. Thurow (Eds.), *The a priori in philosophy* (pp. 53–68). Oxford University Press.

Hughes, R. (1997). Models and representation. *Philosophy of Science, 64*, S325-336.

Isaac, A. (2013). Modeling without representation. *Synthese, 190*(16), 3611–23.

Knott, C. G. (1911). *Life and scientific work of Peter Guthrie Tait* (Vol. 1). Cambridge University Press.

Kuhn, T. S. (1977). A function for thought experiments. *The essential tension: Selected studies in scientific tradition and change* (pp. 240–265). University of Chicago Press.

Mach, E. (1896). *Über Gedankenexperimente. Zeitschrift für Physikalische Chemie Unterrichten 10: 446-457* (W. O. Price & S. Krimsky Eng. Trans.). On Thought Experiments (1973), *Philosophical Forum* 4, 3.

Mach, E. (1919). *The science of mechanics* (Thomas J. MacCormack Eng. Trans.). The Open Court Publishing.

Markie, P. (2021). Rationalism vs. empiricism. The Stanford Encyclopedia of Philosophy.

Maxwell, J. C. (1871). *The theory of heat*. Longmans Green and Co.

Mey, T. D. (2003). The dual nature view of thought experiments. *Philosophica, 72*, 61–78.

Meynell, L. (2014). Imagination and insight: A new account of the content of thought experiments. *Synthese, 191*(17), 4149–4168.

Miščević, N. (1992). Mental models and thought experiments. *International Studies in the Philosophy of Science, 6*(3), 215–226.

Murphy, A. (2020). Thought experiments and the scientific imagination. *Ph.D. Dissertation. University of Leeds*.

Murphy, A. M. L. (2022). Imagination in science. Philosophy. *Compass, 17*(6), e12836.

Nersessian, N.J. (1992) In the theoretician's laboratory: Thought experimenting as mental modeling. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* volume 2 (pp. 291–301).

Nersessian, N. J. (2007). Thought experimenting as mental modeling: Empiricism without logic. *Croatian Journal of Philosophy, 7*(20), 125–161.

Nersessian, N. J. (2018). Cognitive science mental modeling, and thought experiments. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 309–326). Routledge.

Nguyen, J. (2020). It's not a game: Accurate representation with toy models. *The British Journal for the Philosophy of Science, 71*(3), 1013–1041.

Norton, J. D. (1985). What was Einstein's principle of equivalence? *Studies in History and Philosophy of Science, 16*(3), 203–246.

Norton, J. D. (1991). Thought experiments in Einstein's work. In T. Horowitz & G. J. Massey (Eds.), *Thought Experiments in Science and Philosophy* (pp. 129–148). Rowman & Littlefield.

Norton, J. D. (1996). Are thought experiments just what you thought? *Canadian Journal of Philosophy, 26*(3), 333–366.

Norton, J. D. (2004). On thought experiments: Is there more to the argument? *Philosophy of Science, 71*(5), 1139–1151.

Norton, J. D. (2004). Why thought experiments do not transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in the philosophy of science* (pp. 44–66). Blackwell.

Norton, J. D. (2013). Chasing the light: Einstein's most famous thought experiment. In M. Frappier, L. Meynell, & J. R. Brown (Eds.), *Thought experiments in philosophy, science, and the arts* (pp. 123–140). Routledge.

Norton, J. D. (2018). Maxwell's demon does not compute. In M. E. Cuffaro & S. C. Fletcher (Eds.), *Physical perspectives on computation, computational perspectives on physics* (pp. 240–256). Cambridge University Press.

Norton, J. D. (2021). *The material theory of induction*. University of Calgary Press.

Perini, L. (2010). Scientific representation and the semiotics of pictures. In P. Magnus & J. Busch (Eds.), *New waves in philosophy of science* (pp. 131–154). Palgrave McMillan.

Perini, L. (2013). Diagrams in biology. *Knowledge Engineering Review, 28*(3), 273–286.

Radder, H. (1996). *In and about the world: Philosophical studies of science and technology*. State University of New York Press.

Rescorla, M. (2009). Cognitive maps and the language of thought. *British Journal for the Philosophy of Science, 2*(60), 377–407.

Roussos, J. (2020), Modelling in moral philosophy. Unpublished manuscript.

Salis, F. (2016). The nature of model-world comparison. *The Monist, 99*(3), 243–259.

Salis, F., & Frigg, R. (2020). Capturing the scientific imagination. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination* (pp. 17–50). Oxford University Press.

Salis, F., Frigg, R., & Nguyen, J. (2020). Models and denotation. In J. Falguera & C. Martínez-Vidal (Eds.), *Abstract objects* (pp. 197–219). Oxford University Press.

Schabas, M. (2018). Thought experiments in economics. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 171–182). Routledge.

Schlaepfer, G., & Weber, M. (2018). Thought experiments in biology. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 243–254). Routledge.

Shinod, N. (2017). Why thought experiments do have a life of their own: Defending the autonomy of thought experimentation method. *Journal of Indian Council of Philosophical Research, 34*(1), 75–98.

Sorensen, R. A. (1998). *Thought experiments*. Oxford University Press.

Starikova, I., & Giaquinto, M. (2018). Thought experiments in mathematics. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge companion to thought experiments* (pp. 257–278). Routledge.

Strand, R., Fjelland, R., & Flatmark, T. (1996). In vivo interpretation of in vitro effect studies. *Acta Biotheoretica, 44*(1), 1–21.

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., Fehige, Y., & Brown, J. R. (2018). *The Routledge companion to thought experiments*. Routledge.

Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science, 17*(3), 225–244.

Tan, P. (2021). Inconsistent idealizations and inferentialism about scientific representation. *Studies in History and Philosophy of Science Part A, 89A*, 11–18.

Thoma, J. (2016). On the hidden thought experiments of economic theory. *Philosophy of the Social Sciences, 46*(2), 129–146.

Thomasson, A. (2020). If models were fictions then what would they be? In A. Levy & P. Godfrey-Smith (Eds.), *In the scientific imagination* (pp. 51–74). Oxford University Press.

Todd, C. (2020). Imagination aesthetic feelings, and scientific reasoning. In M. Ivanova & S. French (Eds.), *The aesthetics of science* (pp. 63–85). Routledge.

Walton, K. L. (1990). Mimesis *as make-believe: On the foundations of the representational arts*. Harvard University Press.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wilson, J. (2016). VII-Internal and external validity in thought experiments. *Proceedings of the Aristotelian Society, 116*(2), 127–152.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*(6), 981–1022.