Synthese (2021) 199:6119–6147
https://doi.org/10.1007/s11229-021-03063-1

Probing theoretical statements with thought experiments

Rawad El Skaf[1]

Received: 16 July 2020 / Accepted: 28 January 2021 / Published online: 18 February 2021

Abstract
Many thought experiments (TEs) are used to probe theoretical statements. One crucial strategy for doing this, or so I will argue, is the following. A TE reveals an inconsistency in part of our previously held, sometimes empirically well-established, theoretical statements. A TEer or her critic then proposes a resolution in the form of a conjecture, a hypothesis that merits further investigation. To explore this characterisation of the epistemic function of such TEs, I clarify the nature of the inconsistencies revealed by TEs, and how TEs reveal and resolve them. I argue that this can be done without settling the question of which cognitive processes are involved in performing a TE; be they propositional or non-propositional. The upshot is that TEs' reliability, like real experiments, is to be found, in part, in their replicability by the epistemic community, not in their cognitive underpinnings.

Keywords Thought experiments in physics · Inconsistency revealers and resolvers function · Internal vs. external inconsistency · Norton's elimination thesis · TEs' common structure

1 Introduction

Five decades following Kuhn's (1964) question "[h]ow […] relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to new understanding of nature?" (p. 241), scientific thought experiments (TEs) remain a hot topic in philosophy of science. To answer Kuhn's epistemic question, I claim that we should analyse TEs along three[1] interrelated dimensions: their *form*, their

[1] This is not new in the literature. For instance, Meynell (2014) adopts the same strategy and claims that "a complete method for epistemically evaluating any given TE will be a two-step process, shaped by the distinction […] between the *content* and *epistemological function* of TEs" (my emphasis, p. 4163).

✉ Rawad El Skaf
rawadskaff@gmail.com

[1] Department of Philosophy (KGW), Paris-Lodron-Universität Salzburg, Universität Salzburg, Kapitelgasse 4-6, 5020 Salzburg, Austria

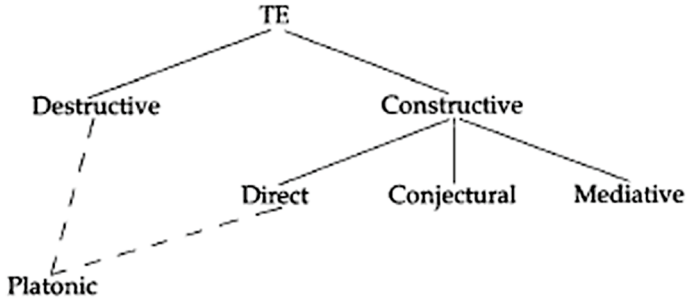

Fig. 1 Brown's taxonomy (2011, p. 33)

content, and their *epistemic function*. Let me start with the third one, the epistemic function of TEs.

In his influential book on the subject (1991, 2011 for the second edition), Brown proposes the following taxonomy of scientific TEs. Basically, they break into two general kinds: *destructive* and *constructive* (Fig. 1). Destructive TEs aim at criticizing a theory, either by revealing an inconsistency (e.g., Galileo's falling bodies) or by showing the theory "to be at odds, not with itself, but with very basic common sense." (2011, p. 35, e.g., Schrodinger's cat). For constructive TEs, Brown distinguishes three subtypes: *mediative*, *conjectural*, and *direct*. A mediative TE (e.g., Maxwell's demon) "facilitates a conclusion drawn from a specific, well-articulated theory" (*ibid*). A conjectural TE (e.g., Newton's Bucket) aims to "establish some (thought experimental) phenomenon; we then hypothesize a theory to explain that phenomenon" (p. 38). Finally, direct TEs (e.g., Galileo falling bodies) "resemble mediative thought experiments in that they start with unproblematic (thought-experimental) phenomena, rather than conjectured phenomena [but like conjectural TEs, they] do not *start* from a given well-articulated theory they *end* with one" (pp. 39–40).

Following this division, Brown defends that there "is a *small class* of thought experiments which are simultaneously in the destructive and the constructive camps" (my emphasis, p. 31). These "platonic TEs" are very special in Brown's approach; they transcend empiricism[2]:

> "A *platonic thought experiment* is a single thought experiment which destroys an old or existing theory and simultaneously generates a new one; it is a priori in that it is not based on new empirical evidence nor is it merely logically derived from old data; and it is an advance in that the resulting theory is better than the predecessor theory." (p. 77).

In my view, *many*[3] TEs not only a *small class,* in physics are both destructive and constructive, including many that Brown identifies as mediative TEs.[4] Nevertheless,

[2] The first part of the 2004 collection *Contemporary Debates in Philosophy of Science* is entitled "Do Thought experiments transcend empiricism?", with Brown and Norton providing opposite answers.

[3] Although I believe that my analysis could be generalized to all scientific TEs, at least in physics, it is beyond the scope of this paper to argue for this.

[4] I thank an anonymous reviewer for drawing this parallel.

these TEs do not *conclusively* destroy an existing theory nor provide a priori access to laws of nature, but function as revealers and resolvers of inconsistencies. First a TE *reveals an inconsistency* in part of our previously held, sometimes empirically well-established, theoretical statements. Then the TEer or her critic offers a *resolution* in the form of a *conjecture* that merits further investigation. This resolution is different from case to case. It could simply point out that we should eject something from our theoretical beliefs, however, without providing a rule as to what to give up. It could also lead, for instance, to changing the scope of a theoretical statement or even adding something new to our existing theoretical statements, however only as a hypothesis that merits further consideration. In all these cases, the resolution of an inconsistency revealed is not linear (more on this below, especially in Sect. 7).

Characterizing the epistemic function of TEs as such[5] helps to address questions about the form and content of TEs. In revealing and resolving inconsistencies, do TEs share a common structure? What is the content of TEs? Most importantly, are TEs reducible to arguments with *eliminable* "particulars" *à la* Norton? To answer these questions, I will focus mainly on Norton's *argument view*. Arguing principally against Brown's platonic account, Norton (since 1991) maintains that TEs do not transcend empiricism. He argues that the only ''non-miraculous'' way to get new information about the world, without resorting to *new empirical data,* is through arguments from premises describing past empirical data about the world. TEs are epistemically dispensable, Norton claims, and a "successful" TE is just a "good" argument with *irrelevant* and *eliminable* "particulars". Norton's argument view thus mainly focuses on the form not the content of TEs, he indeed "tends to stress validity and cogency over soundness" (Stuart 2020, p. 17). But where do we find this neglect of the content of TEs in Norton's view?

Brendel (2018, p. 283) identifies five theses composing Norton's argument view. First, it is mainly in the "Elimination Thesis" (ET), and to a lesser extent in the "Identity Theses", that the *functional* role of the content is poorly considered. Or so I will argue throughout the paper, in particular while analysing the nature of the inconsistencies revealed by TEs in 3. Second, this content is taken to be represented purely propositionally in Norton's "Empirical Psychological Thesis". This is criticized in the literature. For instance, pace Norton's insistence that the actual conduct of a TE is the mental execution of an argument, proponents of the mental model accounts of TEs (see, for example, Gendler 2004; Miščević 1992; Nersessian 1992), argue that a TE's content is not always propositional. As Meynell puts it, "[a]ccording to them, the content of a TE is a mental kind that is not easily reduced to propositions. It is a thought process, a "guided contemplation" (Gendler 1998, p. 414), often with an experiential character, in which the thinker manipulates a mental model in a specified way so as to produce a result" (Meynell 2014, p. 4155).

[5] Many in the literature characterize the function of TEs as such. For instance, the accounts of Kuhn (1964) and Gendler (1998) appraise TEs as tools aimed at testing conceptual frameworks or theories by applying parts of them (usually, central concepts) to new and relevantly unusual fictional scenarios and looking for contradictions. I thank an anonymous reviewer for drawing this parallel.

However, seeing that "the sciences of mind and cognitive science in particular are not only young but highly contentious" (Meynell 2018, p. 499), it seems that this debate will not be scientifically settled, at least, not soon. In addition, these cognitive processes are case, subject and context dependent, as I will argue in Sect. 4. I shall thus endeavour to show throughout the paper that we could analyse TEs and assess their reliability, while remaining neutral as to the nature of the cognitive processes underlying their performance; be they propositional or non-propositional. The upshot is that TEs' reliability lies, in part, in their replicability by the epistemic community, and this can be investigated without reference to their cognitive underpinnings. This said, it remains interesting to investigate this latter subject separately and even draw consequences for TEs (see, for example, Stuart's (2019) analysis of imagination and its implication for several accounts of TEs). This investigation could even complement the approach defended in this paper.

The plan of this paper is as follows. In Sect. 2, I set the stage with a well-known example, i.e., Einstein/Bohr photon-box debate. In 3, I distinguish between two readings of Norton's ET, and argue against the stronger reading by drawing on Krimsky's (1973) distinction between internal and external inconsistencies. In Sect. 4, I describe a common structure for inconsistency revealing/resolving TEs. In Sects. 5 and 6 I provide two illustrations: Einstein's proto EPR photon box, and Einstein, Tolman, and Podolsky's (ETP) photon box. In Sect. 7, I locate the reliability of TEs in their replicability, and formulate, following scientific practice, five strategies for the critic of a TE.

2 Setting the stage: Einstein/Bohr photon-box TE

Let me begin by setting the stage with a well-known and widely discussed TE. The aim in this section is not to introduce something new to the literature, but to analyse the nature of the inconsistencies revealed by TEs in Sect. 3. According to Bohr's (1949) recollection of his informal discussion[6] with Einstein during the 1930 Solvay conference:

> "Einstein proposed the device indicated in (Fig. 2) consisting of a box with a hole in its side, which could be opened or closed by a shutter moved by means of a clock-work within the box" (Bohr 1949, p. 225).

You start by weighing the box, then open the door for a short time in which a single photon can escape from the box, and finally re-weigh the box. The clock gives us the time of passage of a photon, the balance gives us the difference in the mass

[6] The photon-box TE was never published by Einstein and his discussions with Bohr during the Solvay conference were not transcribed. This results in a historical debate concerning Einstein's original aim. While Jammer (1974) agrees with Bohr reading, Howard (1990) for instance defends that Einstein's never sought to refute Heisenberg's uncertainty principle during Solvay's conference. Howard convincingly shows that the historical record shows that Einstein aimed from the beginning to argue against QM's completeness (see proto-EPR photon-box in Sect. 5). For the purpose of my argument in this section, I will treat Einstein's TE as understood by Bohr.

of the box before and after the passage of a photon. By using Einstein's equation $E = mc^2$ we can calculate the energy difference. This energy difference, in accordance with the principle of conservation of energy, would be the energy of the emitted photon. Therefore, one could in principle simultaneously determine the time of escape of a single photon and its energy with an arbitrary degree of precision.

Now, Heisenberg's uncertainty principle says that the simultaneous measurement of two conjugate variables (such as momentum-position or time-energy) for a given particle, results in a limitation of the accuracy of each of these measures. Namely, for the energy and time, the better the accuracy of the energy measurement, the less accurate the measurement of the time will be, and vice versa. The above result is thus.

> "in definite contradiction to the reciprocal indeterminacy of time and energy quantities in quantum mechanics [QM]" (*Ibid*., p. 226).

The story according to Bohr ended with his triumph; he found a flaw in Einstein's scenario the day after that famous discussion. The flaw for Bohr was to be found in the technical and theoretical details[7] of the experimental arrangement of the photon-box:

> "In fact, in the consideration of the problem, […] it was essential to take into account the relationship between the rate of a clock and its position in a gravitational field well known from the red-shift […]. Our discussion concentrated on the possible application of an apparatus incorporating Einstein's device and drawn in Fig. 3" (Bohr 1949, p. 226).

Bohr started by describing Einstein's scenario in more detail, in particular focusing on the weighing procedure:

> "The box, […], is suspended in a spring-balance and is furnished with a pointer to read its position on a scale fixed to the balance support. The weighing of the box may thus be performed with any given accuracy" (p. 227).

Then Bohr argued that this weighing procedure will require the box, and thus the clock, to move in a gravitational field, and.

> "according to general relativity theory [GR], a clock, when displaced in the direction of the gravitational force […] will change its rate." (p. 227).

That is the time measurement by a clock moving in a gravitational field should be given by GR, and not a classical space–time theory, as Einstein seemed to be suggesting. Which led Bohr to reverse Einstein's output and conclude that.

> "a use of the apparatus as a means of accurately measuring the energy of the photon will prevent us from controlling the moment of its escape." (p. 228).

[7] *Cf*. Bishop (1999) for an analysis of Bohr's reply with the theoretical assumptions and mathematical derivation involved.

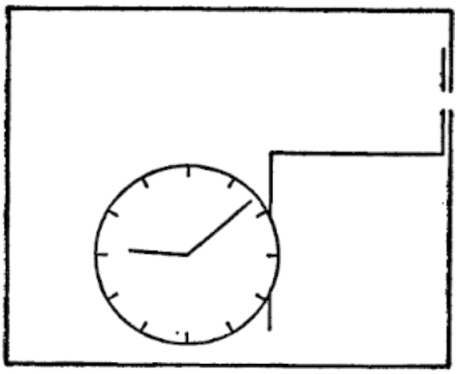

FIG. 7

Fig. 2 Einstein's photon-box (Ibid)

Several morals could be drawn from this episode. Here are two that will be relevant to my analysis in what follows. Einstein's aim (again according to Bohr's reading, see ft.6 and proto-EPR in Sect. 5) was to reveal an inconsistency. Bohr resolved it by better-describing, both technically and theoretically, Einstein's experimental arrangement. We should thus analyse how a TE reveals an inconsistency, its nature, and the content of a TE. For this, I will analyse Norton's different theses, mainly the ET, partly by drawing on Krimsky's distinction between internal and external inconsistency.

3 On the nature of the inconsistencies revealed by TE: Norton's ET and Krimsky's internal vs. external inconsistencies

3.1 Two readings of Norton's ET

In his first paper on TEs, Norton characterises TEs as follows:

> "Thought experiments are arguments which:
> (i) posit hypothetical or counterfactual states of affairs, and.
> (ii) invoke particulars irrelevant to the generality of the conclusion." (p. 129).

Condition (i) gives TEs their "thought-like character", otherwise they "would be the description of real experiments or states of affairs" (*ibid*). Condition (ii) gives them their experiment-like character. One could wonder what Norton means by "irrelevant particulars" in condition (ii). In explaining this, Norton writes that.

> "*The presence of these particulars is what makes thought experiments experiment-like*. Thus, in one version of the thought experiment in which Einstein sought to demonstrate that the effects of acceleration mimic those of gravitation, he asked us to imagine a physicist-observer who has been drugged and reawakens closed up inside a box (Einstein 1912, pp. 1254–55). *That there is an observer, that the observer is a physicist, that the physicist has been*

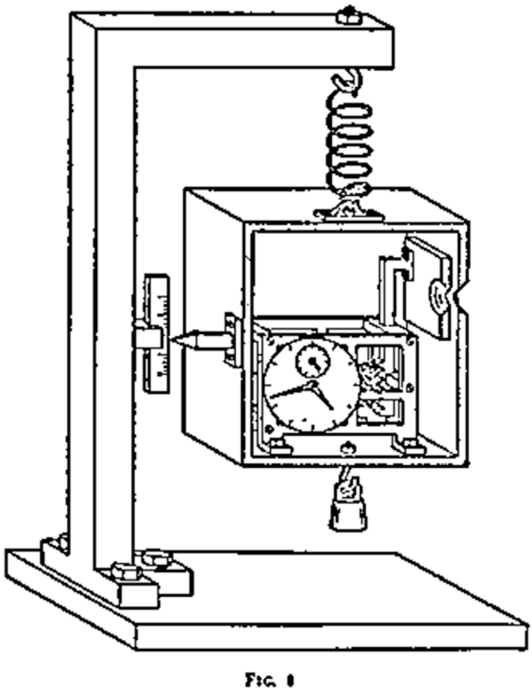

Fig. 3 Bohr's better description of Einstein's photon-box (p. 227)

> *drugged, that he is enclosed within a box*—all these are particulars which are irrelevant to the generality of the conclusion which Einstein seeks to draw. *Without particulars such as these, however, thought experiments would not have their experimental appearance*" (My emphasis, p. 130).

This is by any standard a broad notion of particulars. Any detail that we imagine in the scenario of a TE that gives the TE its "experimental appearance" is counted as a particular. But for Norton's purposes, the notion of "particular" could remain vague for two reasons. First, as Norton explicitly claims, conditions (i) and (ii) above only provides necessary, but *not sufficient* conditions for something to be a TE. From this he adds that.

> "to recover sufficient conditions for a thought experiment from the characterization, *the nature of the particulars in (ii)* would have to be specified more closely. They must be of a type sufficient to *guarantee the appropriate experimental character* to the argument." (My emphasis, p. 130).

Unfortunately, Norton does not elaborate more on this. Instead, and this brings me to the second reason why Norton could remain vague concerning the nature of particulars, Norton goes on and claim that particulars are eliminable:

> *"The elimination thesis* [ET]. Thought experiments are arguments which contain particulars irrelevant to the generality of the conclusion. Thus any conclusion reached by a good thought experiment will also be demonstrable by

an argument *which does not contain* these particulars and *therefore is not a thought experiment*". (My emphasis, Norton 1991, p. 131).

There are two steps here. First a TE is reconstructed as an argument that refers to particulars, let us call this a 'TE-argument'. Second, according to the ET, we could, in principle, be able to transform the TE-argument into a 'non-TE-argument', that is an argument that does not refer to particulars. As some have noted (see, for example, Gendler (1998), Salis and Frigg (2020), Buzzoni (2008) and Stuart (2016)), the ET has two readings, a weaker and a stronger one. Some of these authors interpret Norton's ET in its strong reading (Gendler and Buzzoni) while others (Stuart) take Norton to defend a weaker version. So, which is it? I will argue that Norton is bound to adopt the weaker reading.

(1) *The weaker reading:* This reading claims that we can eliminate *some* details in a given scenario and change others. That is, some of the experimental details are irrelevant to the generality of the *conclusion* and a similar experimental arrangement could provide the same general conclusion. Put differently, the weaker reading is about the *product* of the TE, the conclusion. It says that it is particular-free. It is however not about the *process* that leads to that conclusion. The ET in its weaker version is not saying that the premises that *do* refer to experimental details play no role in TEs, but just that the conclusion is general and particular-free. This is what Norton seems to be defending when reconstructing TEs into TE-arguments. When looking at Norton several reconstructions, we *always* find premises referring to the experimental arrangement. Also, when Norton defends this thesis in the case of deductive TEs, he writes "[t]he particulars might be involved in a counter-example to a universally quantified assertion through which contradiction the conclusion follows" (1991, p. 131). Finally, in 2004, Norton seems to be adopting this weaker reading by claiming that "if all that is required to be experiment-like is that the thought experiment *describe an imaginary experiment and even trace its execution*, then that can be done by an argument." (My emphasis, 2004, p. 62). I will come back to this latter claim later. What is relevant for describing an imaginary experimental arrangement, tracing its execution and interpreting its result will prove to be central to my analysis of TEs. But first, let me analyse the strong reading, just to argue that it could not be defended, even by Norton.

(2) *The stronger reading:* This reading claims that TEs could in principle, even if it is difficult in-practice, be replaced by non-TE-arguments. That is *all* particulars are eliminated in the non-TE-argument that should replace the TE or the TE-argument. Put differently, this reading is not only about the product of a TE, the conclusion, but also about the *process* that leads to the conclusion. It claims that the particulars play no epistemic role in the argument, and the TE could be replaced by a non-TE-argument. Buzzoni for instance took Norton to be defending this reading when he writes that TEs "stripped of any reference to concrete experimental situations, are confined to a domain of purely theoretical statements and demonstrative connections" (2008, p. 67). We find this interpretation in

Norton's papers explicitly when defining the ET as quoted above. Also, at the end of his analysis of what he labels as deductive TEs, Norton concludes:

> "[a deductive TE, like Einstein's black body radiation TE] conclusion follows from [its] premises. Thus we could find another argument *which is not a thought experiment* but which still takes us from premises to the conclusion (the elimination thesis). The argument may well even be a *reductio* argument, but *not one of an experimental character*. But I think it is clear that such an alternative argument would be difficult to find because of the great complexity of [some cases]". (My emphasis, Norton 1991, p. 134).

To better analyse this stronger reading, let me get back to Einstein and Bohr's debate about the Solvay's photon box TE. This episode shows that *not all* particulars, as in experimental details that give the TE its experimental character, are eliminable. Why? The answer lies in Krimsky's (1973) notions of "internal" and "external" inconsistencies, which reflects both the strong and weak readings of Norton's ET, respectively.

3.2 Krimsky's internal vs. external inconsistencies

Krimsky's notions of *"internal"* and *"external"* inconsistencies are developed as a reply to Popper's analysis of the Einstein/Bohr Solvay photon-box episode. Briefly, Popper rejects Bohr's appeal to a "second" theory (i.e., GR) to save a "first" theory (i.e., QM) from the inconsistency revealed by Einstein's TE. Popper argues that this move "amounts to the strange assertion that [QM] contradicts Newton's gravitational theory, and further to the still stranger assertion that the validity of [GR] (or at least the characteristic formulae used, which are part of the theory of the gravitational field) can be derived from [QM]" (Popper 1959, p. 470).

Krimsky's analysis began by clarifying the nature of the inconsistency revealed by the TE. For that, he introduces two notions of inconsistencies:

(1) *Internal inconsistency:* a set of theoretical statements is "*internally* inconsistent" if "we have a purely formal or logical inconsistency derived exclusively from the axiomatics of the theories". (p. 330)
(2) *External inconsistency:* a set of theoretical statements is "*externally* inconsistent" when it "is applied to an experimental arrangement whereupon a statement is implied which is logically inconsistent with one or more statements". (p. 329)

Then Krimsky underlines the differences between the two types in terms of the nature of the procedure employed:

> "The procedure for eliciting the [external] inconsistency is not strictly logical. [It consists in the] *application of theoretical principles to an experimental arrangement* […]. These are theoretical rather than strictly logical procedures." (my emphasis, p. 329).

This brings Krimsky to conclude, *pace* Popper, that Bohr's appeal to GR is perfectly acceptable. There is nothing peculiar about Popper's remark since "it is not at all strange that [QM] and Gn [Newton's gravitational theory] be externally inconsistent with respect to some experimental arrangement" (*Ibid*). While if we analyse the TE as aiming at revealing an internal inconsistency—which would be the case under the stronger reading of the ET, where the experimental details are eliminated in order to transform the TE into a non-TE-argument—then Popper would be justified in claiming that this amounts to the strange assertion that QM and Gn, as a set of theoretical statements, are internally inconsistent.

And Norton agrees with this reading. In replying to Bishop's (1999) criticism against his identity theses,[8] Norton writes:

> "Einstein and Bohr *do* have two different, but similar, thought experiments; and they correspond to two different, but similar, arguments. We can convert the two thought experiments into one by ignoring the different spacetimes of each. The different spacetime settings are then responsible for the different outcomes. If that is admissible, then the same stratagem works for the arguments. Ignoring premises pertaining to the spacetime setting, the two arguments proceed from the same experimental premises. They arrive at different results only because of the differences in the premises pertaining to spacetime setting." (2004, p. 64).

This shows that Norton is adopting the weaker reading of the ET, since the two arguments proceed from the same "experimental premises". However, it remains unclear to me how we should understand the idea that we convert the two TEs into one, by ignoring the different spacetimes of each. How does that work and what is left in the content of the resulting common TE? Is a TE's content only about experimental arrangements or is also about the way we trace its execution and interpret its result? If the latter, then the space–time setting is crucial. Norton's move is even more perplexing, seeing that experimental details, not theoretical statements, are usually ignored in Norton's ET!

Be that as it may, here is how I think this episode is best understood. Einstein imagined an experimental arrangement and traced its execution (implicitly or explicitly) using Gn. He arrived at an outcome which contradicted Heisenberg's principle. An inconsistency is thus revealed. Bohr came the next day and proposed a resolution. He argued that a "pseudorealistic" experimental arrangement should be traced using GR, not Gn. If we do this, then the outcome that contradicted Heisenberg's principle no longer follows. It seems we have here a single TE: the original Einsteinian version revealed an inconsistency, while Bohr's reply proposed a resolution, in

[8] Briefly Bishop argues that TEs could not be identified as arguments since we have one TE but two arguments See Bishop (1999), Norton's reply (2004, p. 64) and Brendel (2018, pp. 284–285) concerning this issue. Also, see Bokulich and Frappier (2018) for a general discussion of the identity criterion in several epistemic accounts of TEs.

the form of conjecture.[9] This interpretation captures what happens in several TEs, ETP in Sect. 6 is a case in point.

Let me consider another example to make my point apparent. Vickers provides a nice case study from Pauli, aimed at revealing an inconsistency in Bohr's theory of the atom and old quantum theory:

> "First Pauli considered the hydrogen atom with electric and magnetic external fields coming from the same direction as in Fig. 4(a). This gave certain allowed orbits, decided by giving the appropriate values for the quantum numbers n, k, m, and s in the quantum condition [i.e., only certain orbits are possible]. Then Pauli considered an adiabatic change to the system, rotating the electric and magnetic fields in opposite directions as in Figure 4(b). *Following the rules of the adiabatic principle, it turned out that by doing this a system could be achieved theoretically* for which the magnetic quantum number m was equal to zero. But […] $m \neq 0$ was stipulated as a necessary feature of the quantum condition." (My emphasis, Vickers 2013, pp. 65–66).

This is clearly a TE and is aimed at revealing an external inconsistency in Bohr's old quantum theory. Pauli is applying general principles—the quantum condition and the adiabatic principle—to describe an experimental arrangement and trace its execution—a hydrogen atom with the E and B fields. The quantum condition allows certain orbits to the initial system, "decided by giving the appropriate values for the quantum numbers n, k, m, and s". This initial system is then rotated "following the rules of the adiabatic principle". By doing this we get a theoretical system with a magnetic quantum number $m = 0$. This outcome is directly interpreted as contradicting a necessary feature of the quantum condition, viz. $m \neq 0$.

Put differently using Krimsky's terminology, we have here an external inconsistency revealed between two theoretical statements.This is the case seeing that "there are literally hundreds of ways in which one might apply electric and magnetic fields to a hydrogen atom and then consider a possible adiabatic transformation of those fields [without having $m = 0$]. It may be that the particular way Pauli did it was *the only way* in which to derive a contradiction" (My emphasis, Vickers 2013, pp. 68–69). In addition, these general principles in and by themselves *do not directly contradict*. To see this, contrast Pauli's TE with other attempts to reveal an inconsistency in Bohr's theory.[10] For instance, when it is argued that the quantum condition—which posit only discrete energy levels—directly contradicts theoretical statements from classical electrodynamics—which posit a continuity of energy levels.

[9] Many versions of the photon-box TE are still being proposed, each varying in the details of the experimental setup and the theoretical statements. Stuart (2016) nicely summarizes this by noting that the photon-box TE "continues to be debated in the same context as it was originally presented, namely, concerning whether and how the uncertainty principle maintains itself in the face of certain possible experimental arrangements (see Hilgevoord 1998; Hnizdo 2002; Kudaka and Matsumoto 1999; Treder 1970)." (p. 28).

[10] *Cf.* Vickers chapter 3 for an analysis of these attempts. Vickers concludes that Pauli's TE reveals the most serious inconsistency.

It is clear that these generally formulated theoretical statements directly contradict, they are internally inconsistent.

What about the resolution? From the inconsistency revealed Pauli concluded that "[a]n escape […] can be achieved only by a radical change in the foundation of the theory" (Pauli 1926: 163–4, quoted from Vickers p. 66). These resolutions, whatever they may have been, were not historically pursued. As Vickers notes "the question […] which of the propositions in the Pauli inconsistency can be ejected with minimal damage [didn't] really arise in a serious way in the relevant scientific history, because by 1926 Heisenberg's matrix mechanics and Schrödinger's wave mechanics were changing the scientific landscape dramatically" (Vickers, p. 70).

3.3 Transforming an external to an internal inconsistency

More generally, the stronger reading could only be defended if we can *always* be assured that we could transform an external inconsistency into an internal one. Norton clearly thinks we can. Recall his conclusion above "[a TE's] conclusion follows from [its] premises. *Thus* we could find another argument *which is not a thought experiment* but which still takes us from premises to the conclusion (the elimination thesis)."

The problem here is that Norton does not provide any argument to establish that such a transformation is always possible, even *in-principle*.[11] For a TE (or a TE-argument) to be transformed into a non-TE-argument according to the strong reading of the ET, what is needed is an independent argument. An argument that somehow takes us from the observation that an external inconsistency is revealed by a TE (the a "[TE's] conclusion follows from [its] premises" part of Norton's claim) to the conclusion that the inconsistency is somehow contained in these general statements (the "*thus* we could find another argument *which is not a thought experiment"* part). It could be that this transformation is possible in some cases but it is clear that it will not be possible for all cases, at least before the end of inquiry. For instance, to do this in the photon box episode, we need following Krimsky a new theory:

> "Since the photon box experiment shows that quantum mechanics is rescued from external inconsistency by [GR] we may wish to define a new quantum mechanics which includes Einstein's general theory of gravitation, [QM' = (QM and GR)]. Now we can argue, quite reasonably, that [QM]' is internally inconsistent with Gn (since Gn and [GR] are internally inconsistent)." (*Ibid*).

If this is correct, then it seems too much to ask. We need a fundamental unifying theory not even available today; nearly a century after the photon-box TE. When, and if, we have such a theory, we could argue that this new QM, QM', is internally

[11] Things become a lot more complicated, as Norton is fully aware, if we adopt an *in-practice* approach and inquire whether it is possible for scientists to derive an internal inconsistency directly from the general theoretical statements involved in the TE, *with the theoretical/computational/mathematical knowledge at their disposal at the time the TE was conceived and used.*

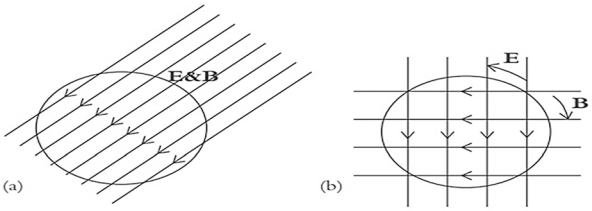

Figure 3.1 (a) Electric and magnetic fields coming at a hydrogen atom from the same direction; (b) the directions of the electric and magnetic fields are turned infinitely slowly and smoothly (as mathematically conceived) in opposite directions until they cross each other at right angles

Fig. 4 Pauli's TE (ibid)

inconsistent with Gn. But this is too much speculation on how future physics would look like. Also, the TE would have served its purpose by then. I will not follow this further; the burden of proof is on Norton if he wishes to defend that *all* external inconsistencies are transformable to internal ones, which I doubt he wishes to do.

Indeed, considering all the arguments in this section, Norton's ET in its strong reading should be rejected. I also think that Norton's argument view comes out stronger with the weaker reading; I see no reason to eliminate the "experimental character" of TEs in an epistemic account of thought *experiments.* Even Norton's "Reliability Thesis" remains consistent with the weaker reading. This latter just "connects the reliability of a [TE] to the justifiedness of the conclusion of the argument into which the [TE] is reconstructed." (Brendel 2018, p. 286). Nothing in the reliability thesis requires that such an argument should be a non-TE-argument. I will come back to the reliability of TEs in Sect. 7. Granted that *not all* particulars are eliminable, then what are their functions? There are several,[12] one important function is to group theoretical statements together in order to, as Norton puts it, describe, trace the execution of an experimental arrangement and interpret its result.

3.4 Grouping theoretical statements

In analysing inconsistencies in science, Vickers defends that "the inconsistency of a set of propositions will only be interesting and/or important if there is some point to grouping those propositions together" (2013, p. 31). TEs are the tool *par excellence* for grouping theoretical statements in a single scenario of an experimental character. This will become more apparent in the case of the proto-EPR TE (see Sect. 5). It is also the case in both Pauli's TE and Einstein/Bohr debate. This latter shows that the grouping of theoretical statements initially proposed could be challenged, as Bohr did (more on this in Sect. 7.2).

[12] Some details serve to single out a specific causal factor (e.g., Galileo's falling bodies, *cf.* El Skaf 2018). Other details, or lack of details serve to abstract the theoretical and technical description of some processes (e.g., Maxwell's demon, see Sect. 4 and El Skaf 2017).

More generally, grouping theoretical statements in a single scenario is crucial for science and its progress. Recently, many TEs have been used in black hole physics (since the 1970′s). In these TEs, statements from QM, GR and thermodynamics, our best current theories, are grouped together. Their aim is precisely to reveal an inconsistency between these well-established theoretical statements; the "information paradox". Its resolutions are analysed in different research programs, spanning from QM to GR. These programs diverge, in part, following which theoretical statement scientists are willing modify or even completely reject. These TEs are an important tool in contemporary foundational (philosophy of) physics, seeing that the object of inquiry is *inaccessible*[13]—which renders any direct investigation impossible, at least for the time being. In addition, these theoretical statements come from theories with different objects and scales of applicability and thus are not easily grouped together.

Granted that some experimental details are ineliminable, we now need an account of TEs which appraises both their theoretical and experimental character. In the remainder of the paper I shall propose such an account.

4 TEs' common structure

Having delimited the epistemic function and analysed both the nature of the inconsistency revealed and the functional role of the content, I will now consider the form of TEs. The TEs I am highlighting in this paper, those that reveal and resolve inconsistencies, have the following common structure. This structure aims to be general enough to accommodate many TEs, with different degrees of theoretical description. Nevertheless, specific enough to appraises both TEs' theoretical and experimental character. Finally, it is neutral as to the nature of the cognitive processes underlying TEs' performances; be they propositional or non-propositional.

1. *Target Theoretical Question(s):* The scientist identifies a target question(s) and uses a TE to answer it/them. For that, he/she imagines a scenario (2), unfolds it (3), arrives at a result, the OU, (4), which, when interpreted reveals an inconsistency (5) which calls for a resolution (step 6)
2. *Scenario:* Contains a more-or-less well-described experimental arrangement and is theoretically delimited. More precisely, the scenario of a TE is composed of the following elements:

    (a) Theoretical/empirical statements (quantified or not)[14]
    (b) Experimental arrangement, involving objects and things that happen to or are performed by them.

[13] Black holes and the theoretical physics involved are a very interesting subjects to inquire further into TEs. More generally to analyse different scientific tools beyond direct empirical experiments, such as Analogue Experiments. That could be interesting seeing that TEs aim at non-empirical disconfirmation, while Analogue experiments aim at non-empirical confirmation. But that is another story!

[14] I will drop the empirical and refer to these statements as theoretical in the remainder of the paper.

(c) Description of the behaviour of the theoretically under-described parts of the experimental arrangement
(d) Idealisations and abstractions

3. *Unfolding of the scenario*[15]*:* We apply the theoretical statements in 2 (a) and the description of the behaviour of the theoretically under-described parts of the experimental arrangements in 2 (c), to describe and trace the execution of the experimental arrangement in 2 (b).
4. *Output of the unfolding, the OU:* If the unfolding of the scenario is correctly done (more on this in Sect. 7), we obtain a proposition as an output, the OU. It is crucial to distinguish the result of such unfolding, the OU, from the conclusions of the TE, steps 5 and 6 below.
5. *Inconsistency revealed*: It is by interpreting the OU with a piece of argument that an inconsistency, real or apparent, is revealed.
6. *Inconsistency resolved*: The TEer offers a way out of the inconsistency revealed in the form of conjectures, a hypothesis to be further explored and tested by future theoretical developments and empirical confirmation.

Several things need clarification. First the unfolding in step 3 is intended to capture the idea that tracing the execution of the imagined experiment is mainly, but *not entirely*, theoretically driven. As well as that this tracing (be it completely or partly theoretical), is not cognitively determined.

First, how we apply theoretical statements from 2(a) to describe and trace the execution of the imaginary experimental arrangement is (i) *case*, (ii) *subject* and (iii) *context* dependent. (i) It is case dependent seeing that TEs' scenarios are diverse. We are asked to reason about measuring instruments as well as microscopic, macroscopic, all the way to massive celestial objects such as black holes. Since these are not actual objects present (and sometimes non-existent objects that could not even be present) to sensory experience, tracing *things that happen to or are performed by them* is mainly guided by previously acquired theoretical and empirical knowledge. How we trace things that happen to or are performed by imagined macroscopic falling bodies and hypothetical clocks is quite different from how we do this for quantum particles and black holes. Our intuitions of how the former objects behave is more stable than the latter. We are macroscopic beings that interact with medium size objects all day long, but we have never seen an electron or a black hole. (ii) It is context dependent. When we look at early modern TEs with our current knowledge is surely different than how seventeenth century reader would look at it. (iii) It is subject dependent. While some readers would need to unfold the scenario by explicitly following these theoretical statements (e.g., Bohr's reply to Einstein). Others could effortlessly "see" or "intuit" how to trace the execution of the experimental arrangement without explicitly activating these theoretical statements. These latter function as implicit background knowledge. In all these cases, when assessing TEs,

[15] *Cf.* El Skaf and Imbert (2013) for an analysis of TEs, Computer Simulations and Real Experiments as unfolding scenarios following part of this common structure.

ideally a direct calculation or derivation would form the ultimate justification that the experimental arrangement is correctly described and traced by these theoretical statements. Put differently by Brendel:

> "[T]here are some a posteriori acquired "truths" that function as implicit background knowledge, enabling us to come to a relatively quick decision in the evaluation of a thought experiment. But we can always make these premises explicit by reconstructing the thought experiment as an argument. This is useful, for example, when we are sceptical about the plausibility of a thought experiment." (2004, p. 96).

This means that Norton's reliability thesis could remain the ultimate guide. Without committing us to the ET in its strong reading or to the identity thesis, as I have argued in 3. Neither committing us to the empirical psychological thesis. This latter thesis brings me to my second point: not every detail of the experimental arrangement is, or even could be, theoretically or empirically described and traced. In these TEs, we seem to go beyond 2 (a). This is something that is meant to be captured by 2 (c) in the structure. To see this, consider Maxwell's original demon TE:

> "suppose that […] a vessel [full of gas at equilibrium] is divided into two portions, A and B, by a division in which there is a small hole, and that a *being, who can see the individual molecules*, opens and closes this hole, so as to allow only the swifter molecules to pass from A to B, and only the slower ones to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics." (Maxwell 1871, p. 309).

From this contradiction, Maxwell goes on and defends that the second law should be statistical. There is clearly a lot to say about this TE, here I would like to emphasis the following aspect. Maxwell's demon TE is a canonical TE in which we are presented with a fantastic "demon", "being", "ghost" or "intelligence"; endowed with certain faculties.[16] In Maxwell's own words, his demons as Thomson named them are "very small but lively beings incapable of doing much work but able to open and shut valves which move without friction or inertia" (undated letter to Tait, entitled "nature of demons", quoted from Canales 2020, p. 52). Put differently, a crucial part of the scenario is theoretically under-described, viz. the molecular separation process. We are asked to imagine a demon, a fictive object, and the description of the things that happen to or are performed by it are provided by the author in 2(c), without referring to theoretical statements.[17]

[16] *cf.* Canales 2020 *Bedeviled: A Shadow History of Demons in Science,* for a recent historical analysis of demons in science, mainly in TEs, including Descartes', Laplace's, Maxwell's, Darwin's and Einstein's demons.

[17] The presence of such an under-described mechanism in the TE is however not detrimental for Maxwell's purpose. C*f.* El Skaf 2017 for an analysis of this. Briefly in this paper I argue that under the following conditions the molecular separation process could remain under-described and its theoretical possibility could remain indeterminate: first Maxwell did not totally reject the second law in the TE's conclusion. Second the TE's OU—fast molecules on one side, slow one on the other—is theoretically possible under Maxwell's kinetic theory of gas.

The question now is in imagining such a molecular separation process, are we in line with the argument view, entertaining the *proposition* "imagine/suppose *that* a demon could measure molecules speeds and manipulate a massless door as described". Or in line with the mental model views, *non-propositionally imagining* a demon and then picturing in our "mind's eye" the demon detecting fast and slow molecules and manipulating a massless door so as to separate fast from slow molecules. I claim that neither is important in assessing this TE and a way out of this debate would be to adopt a pluralist approach to cognitive processes. That is to defend in the same spirit as Cooper that "[w]hether the thought experimenter reasons through the situation via manipulating a set of propositions, or a mental picture, […], makes no difference to my account" (Cooper 2005, p. 341). This is the case for describing and tracing both theoretically well-described and under-described parts of the experimental arrangements.

Let me move now to steps 4 to 6. Briefly (and I will come back to thus in Sect. 7) step 4, the OU, is analogous to the empirically obtained observational statements that constitute the result of some real experiment. The OU is propositional and is the piece of non-empirically obtained evidence that is conflicting with one or more theoretical statements. It is by interpreting the OU that the inconsistency is revealed in step 5 and a resolution is proposed in step 6.

This structure is directly applicable to case studies discussed in previous sections. In the following two sections I provide two illustrations: Einstein's proto EPR photon box, and Einstein, Tolman, and Podolsky's (ETP) photon box. For each, I will rely on the narrated scenario, but I will only focus on the three final steps, the remaining parts of the above structure being easily identifiable from each TE's narrated scenario.

5 Einstein's proto-EPR photon-box

Howard (1990) convincingly argues that Einstein did not use the photon-box TE against the consistency of QM, but against its completeness, even during Solvay's meeting. For that he provides several historical sources, most importantly the following, not even a year after Solvay's meeting:

> "He [Einstein] said to me that, for a very long time already, he absolutely no longer doubted the uncertainty relations, and that he thus, e.g., had BY NO MEANS invented the "weighable light-flash box" (let us call it simply L-F-box) "contra uncertainty relation," but for a totally different purpose." (Ehrenfest to Bohr, 9 July 1931, quoted in Howard 1990, p. 97).

The different reason that Ehrenfest is referring to is of course QM's completeness, a kind of proto-EPR TE. In their 1935 article, Einstein,Podolsky and Rosen (EPR) presented the famous argument or TE against QM's completeness. EPR is still largely debated by both scientists and philosophers. Most however agree that EPR is subject to several interpretations. Even Einstein was not happy with the way

it was presented.[18] Here I will thus only focus on Einstein's own proto-EPR photon box TE. Howard (1990) nicely summarizes Einstein's criticism against completeness as follows. It is interesting to quote the full passage:

> "A complete theory assigns one and only one theoretical state to each real state of a physical system.' But in EPR-type experiments involving spatio-temporally separated, but previously interacting systems, A and B, quantum mechanics assigns different theoretical states, different "psi [ψ]-functions," to one and the same real state of A, say, depending upon the kind of measurement we choose to carry out on B. Hence quantum mechanics is incomplete.
> The crucial step in the argument involves the proof that system A possesses one and only one real state. This is held to follow from the conjunction of two principles that I (not Einstein himself) call the locality and separability principles. Separability says that spatio-temporally separated systems possess well-defined real states, such that the joint state of the composite system is wholly determined by these two separate states. Locality says that such a real state is unaffected by events in regions of space–time separated from it by a spacelike interval."

Frist a TE or an argument against completeness of QM must include a proof that a system has one and only one real state. This proof follows from the conjunction of the two principles of separability and locality. What is needed now is to argue that all these theoretical statements could be grouped together and applied to describe, trace the execution of an experimental arrangement and interpret its result. Indeed, Howard directly continues:

> "Einstein argues *that both principles apply to the separated systems* in the EPR-type experiment." (my emphasis, Howard 1990, pp. 64–65).

Let us see how all this is done in Einstein's proto-EPR TE.[19] Howard finds several occurrences of this TE in Einstein's letters (starting in 1932). The most direct one is to be found in Einstein's (1945) letter to Paul S. Epstein, a Cal-Tech physicist. Briefly, an experimenter, with various measuring instruments, rides a moving photon-box which later comes to rest. Now, as before a particle escapes through a shutter with a clock work mechanism. However now, we do not expect to do both measurements, but we.

> "can either [do] a precise measurement of the box's position, and thus a prediction of the exact time when the emitted photon will be received at some distant location S […]. Or the experimenter can make a new measurement of the

[18] Einstein voiced his concerns to Schrödinger, before presenting his own argument against the completeness of quantum mechanics "For reasons of language this [paper] was written by Podolsky after several discussions. Still, it did not come out as well as I had originally wanted; rather, the essential thing was, so to speak, smothered by the formalism [Gelehrsamkeit]." Quoted from Fine 1986 p. 35.

[19] Proto-EPR differs however from EPR in that one of the systems is a macroscopic object, a box, and not an entangled particle as in EPR.

box's recoil momentum, in which case he or she can predict exactly the energy […] of the emitted photon." (Howard 1990, p. 102).

Einstein then interprets the effect of this choice on the fleeing particle as follows:

"As soon as it has left the box B, the light quantum represents a certain "real state of affairs," about whose nature we must seek to construct an interpretation, which is naturally in a certain sense arbitrary.
This interpretation depends essentially upon the question: should we assume that the subsequent measurement we made on B physically influences the fleeing light quantum, that is to say, the "real state of affairs" characterized by the light quantum?
Were that kind of a physical effect from B[Box] on the fleeing light quantum to occur, it would be an action at a distance, that propagates with superluminal velocity. Such an assumption is of course logically possible, but it is so very repugnant to my physical instinct, that I am not in a position to take it seriously.

That is, both separability and locality are used here to interpret the fleeting light quantum. Separability associates with it an independent "real state of affairs" from the moment it leaves the box. While locality negates that the subsequent measurement made on the box could physically influences the fleeing light quantum, seeing that would be an action at a distance. From that Einstein concludes:

Thus I feel myself forced to the view that the real state of affairs corresponding to the light quantum is independent of what is subsequently measured on B. But from that it follows: every characteristic of the light quantum that can be obtained from a subsequent measurement on B exists even if this measurement is not performed. Accordingly, the light quantum has a definite localization and a definite color.
Naturally one cannot do justice to this by means of a wave function […] This is what I mean when I advance the view that quantum mechanics gives an incomplete description of the real state of affairs" (Letter from Einstein to Epstein, quoted in Howard 1990, p. 102–103).

5.1 Einstein's proto-EPR photon-box structure

1. *OU*: We can in principle determine *either* the exact time *or* the exact energy of the emitted photon depending on the measurement done on the box, however after the separation of the particle.
2. *Inconsistency revealed*: The OU is "logically possible", but repugnant to Einstein's "physical instinct". That is, for QM to offer complete description, either locality or separability (or both) should be given up. The TE brings thus to light an inconsistency between locality/separability and completeness.
3. *Inconsistency resolved*: Einstein maintained both locality and separability and concluded that QM must be incomplete. Which means that we can hope to find the theory that would ultimately complete or even replace QM. While Einstein

> sought to build a unified field theory within which the results of the new quantum theory would be derived, many did not agree, and mainstream physics took another direction.

This shows contrary to Brown's platonic account, that EPR or proto-EPR TEs is not a platonic TE. It *did not* "destroy[] the Copenhagen interpretation and establish[] hidden variables" (Brown 1991, p. 77) or even aimed at this but failed as Brown argues. As many have noted (Howard 1990; Fine 1986; Bokulich 2001), EPR-type TEs aimed only at revealing an inconsistency between locality/separability and completeness, and at proposing a resolution as a conjecture. Indeed, Einstein even finishes the above letter by providing different possible resolutions, maintaining QM completeness. These show some of the different research programs available, which were also historically pursued:

> "If one is of the view that a theory of the character of quantum mechanics is definitive for physics, then one must either completely renounce the spatio-temporal localization of the real, or replace the idea of a real state of affairs with the notion of the probabilities for results of all conceivable measurements. I think that this is the view that most physicists currently have in mind. But I do not believe that this will prove to be the correct path for the long run." (*Ibid).*

6 ETP's photon-box

The photon-box was reused in 1931 for a third purpose by Einstein, Tolman, and Podolsky (ETP). In these 2 pages' paper solely dedicated to the TE, ETP describe a new way to use the photon-box apparatus: in this scenario, two particles escape from two different holes opened at the same time by a clockwork mechanism. One particle follows the short path, *SO*, and the other the long path, *SRO*, reflected to *O* by a mirror at *R* (Fig. 5).

As other photon-boxes, ETP involve some measurement instruments:

> "The box is accurately weighted before and after the shutter has opened in order to determine the total energy of the particles which have left, and the observer at *O* is provided with means for observing the arrival of particles, a clock for measuring their time of arrival, and some apparatus for measuring momentum.

ETP continue by describing the measurements performed on the first particle and the box:

> Let us now assume that the observer at *O* measures the momentum of the first particle as it approaches along the path *SO*, and then measures its time of arrival […]. [K]nowing the momentum of the particle in the past, and hence also its past velocity and energy, it would seem possible to calculate the time when the shutter must have been open from the known time of arrival of the

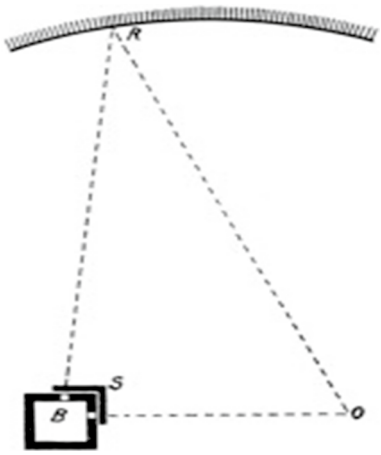

Fig. 5 ETP's photon-box (p. 780)

first particle, and to calculate the energy and velocity of the second particle from the known loss in the energy content of the box when the shutter is opened.

This violates Heisenberg's principle since:

It would then seem possible to predict beforehand both the energy and the time of arrival of the second particle.

Nevertheless, ETP will reject this violation of Heisenberg's principle and defend that the

explanation of the *apparent paradox* must lie in the circumstance that the past motion of the particle *cannot be accurately determined as described*." (My emphasis, ETP, p. 781).

6.1 ETP's structure

1. *OU*: We can simultaneously determine both the energy and the time variables for the second particle by measurements done on the box and the first particle.
2. *Inconsistency revealed*: The OU contradicts Heisenberg's principle.
3. *Inconsistency resolved*: ETP argue that this inconsistency is only apparent since "the particle's past path cannot be accurately determined as described". But for that, it must be concluded that Heisenberg's principal scope should be extended and "must involve an uncertainty in the description of past events" (*ibid).*

First, note the similarity between ETP and Einstein/Bohr Solvay's photon box. Both ETP and Solvay's Einstein reveal the same inconsistency in the OU.[20] Both

[20] Do these two TEs reveal the same inconsistency? In a sense yes, they have the same proposition in the OU. However, we arrive at both OU following a different experimental arrangement. I thank an anonymous reviewer for pointing this.

ETP and Solvay's Bohr proposed a resolution by paying more attention to the tracing of the experimental arrangement. Seeing that we quite naturally consider ETP as a single TE, we should do the same for Solvay's photon-box (as I did in Sect. 3.2). Second, it is interesting to note that ETP's Einstein did not argue against Heisenberg's principle, 1 year after he allegedly attacked it during Solvay. This is more perplexing seeing that OU is the same and ETP's scenario is stronger than that of Solvay's. Stronger in the sense that ETP explicitly configure their experimental arrangement in such a way to block Bohr's appeal to GR:

> "the distance *SO* [is] sufficient so that the rate of the clock at *O is not disturbed by the gravitational effect involved in weighing the box*, and the distance *SRO* being very long in order to permit an accurate reweighing of the box before the arrival of the second particle" (My emphasis, *ibid*).

That is the experimental arrangement is crucial to determining which theoretical statements are relevant, as I have argued in Sect. 3.4. Here it is made explicit that GR is irrelevant, contrary to Bohr's reply during Solvay's meeting. The question now is why didn't ETP argue against Heisenberg's principle? The answer seems to lie not in the TE, but in the confidence scientists accord to a given theoretical statement, here Heisenberg's uncertainty principle. This shows, again, that the inconsistency revealed,[21] by itself, does not point to a *single* resolution. It will not be a stretch to say that the decision is (sometimes) pragmatic. Indeed, instead of defending and extending the scope of Heisenberg's principle, we could easily imagine ETP using their TE to argue against it.

7 The reliability of TEs and five strategies for the critic

7.1 Reliability in replicability

So far, my analysis did not explicitly address the *reliability* of TE. If TEs are not arguments with irrelevant particulars, then why should we trust their conclusions? To assess the reliability of TEs, it is crucial to clearly distinguish between the two conclusions of a given TE, i.e., steps 5 and 6 in the structure (see Sect. 4). While 5 seems the most robust conclusion, 6 is best interpreted as guiding future research programs. In both, TEs share an interesting analogy with real experiments (REs). I will address both. Let me start with 6. Even in light of new unexpected empirical evidence against a theory, the resolution of a tension (an inconsistency) between *new empirical* evidence and theory is messy. As Patton notes in defending the independence of the "context of pursuit"—which she distinguishes in the work of "Laudan's (1977), Nickels's (1981), and Franklin's (1993)" (p. 235, abstract)—from the context of justification, she writes:

[21] Do these two TEs reveal the same inconsistency? In a sense yes, both scenario lead to an output which contradicts Heisenberg's principle. However, we arrive at both OU following a different experimental arrangement. I thank an anonymous reviewer for pointing this.

> "The existence of new evidence in conflict with an accepted theory is not merely an engine of linear change from one theory to the next. It provokes a broad set of questions, which lead the scientific community to make decisions about which theories or hypotheses to pursue, to aid in solving the problem of nonconforming evidence. These decisions about what to pursue are risky – they are not determined fully by the available evidence. […] Ideally, the process of pursuit will yield results that indicate the kinds of solutions to the problem that are available" (2012, p. 245).

TEs play an analogous role. The OU in step 4 is the piece of *non-empirical* evidence and step 5 makes explicit how it conflicts with an accepted theoretical statement. The context of pursuit tells us that the resolution, in step 6, is only conjectural and opens research programs. This nicely captures how some TEers have understood these resolutions. For instance, Einstein's own listing of the different resolutions following the inconsistency revealed in proto-EPR (see Sect. 5 and 7.2.5 below). Which brings me to steps 3, 4 and 5 of the structure in Sect. 4.

Both real and thought experiment's outputs—observational/empirical for the former and the OU for the latter—are better supported when the thought/real experiment is *reperformed*. The parallel between TEs and REs is not new on the literature on TEs (e.g., Gooding 1992; Stuart 2016; Sorensen 1992; Arcangeli 2018). Here I will concentrate on Sören Häggqvist's (2009) modal account of TEs. For Häggqvist, both REs and TEs should not be identified as arguments *à la Norton*. However, both "work only through their connection with arguments" (p. 62). Formally, both TEs and REs are similar, but for the former the argument schema is modal:

> "'C' is the counterfactual scenario described in the thought experiment, 'T' is the theory to be tested, and 'W' is a statement claimed by the thought experimenter to be (i) false in the counterfactual scenario, yet (ii) one to whose truth the theory under testing is committed in that scenario.
>
> $$(\alpha) \qquad \begin{array}{l} \Diamond C \\ T \supset (C \mathrel{\Box\!\!\rightarrow} W) \\ C \mathrel{\Box\!\!\rightarrow} \neg W \\ \hline \neg T \end{array}$$
>
> " (p.63)

Notwithstanding some important differences, especially concerning the modal nature of Häggqvist's account and the non-modal character of my structure, several similarities exist between our accounts. Both identify the main function of many TEs as inconsistency revealers (however, notice in (α) conclusions 5 and 6 are not separated). Both are pluralist relative to the cognitive processes.[22] We both reject the ET in its strong reading (*cf.* Häggqvist 2009, p. 75, ft. 48. Häggqvist

[22] "consider the mechanisms resulting in belief in premises of an argument instantiating (α) when a thought experiment is performed. These may draw on all sorts of cognitive resources: theoretical and other belief, memory, inference, genetically inherited modal expectations, folk physics, and so on. If Nersessian and others are right, they may involve manipulation of mental models. If Brown is right, they may include special faculties of intellectual *Schauung*. If David Chalmers, Frank Jackson, and an earlier time slice of Stephen Yablo are right, modal claims such as the premises of a regimented thought experiment may be justified by appeal to conceivability." p. 72.

however seems to be also rejecting the weaker reading). Both accounts "seem vaguely Popperian in [their] emphasis on *counterinstances* to the target thesis". However, and I agree that "it is a fact that this is what thought experimenters are typically in the business of trying to construct. And this is, in turn, only to be expected. Confirmation is a tricky notion in the best of settings," (p. 64). Indeed, theory choice and confirmation are already tricky, even when new empirical data are generated. As we saw above in Patton's context of pursuit. When dealing with armchair inquiry—which, by definition, do not produce any *new* empirical data—it is hard, at least for an empiricist, to see how TEs could justify new laws of nature or unambiguously decide in favour of one theory and against another.

Finally, both accounts accord some reliability in the replicability of TEs by the scientific community. Häggqvist is however less optimistic. He mentions in closing two features of TEs that differentiate them from REs. The second I will not analyse here, seeing that it is not relevant to my account (it is related to the nested counterfactual conditional, the second premise in (α) which does not appear in my structure). The first is the following:

> "The sheer lack of […] intersubjective agreement in the case of thought experiments indicates that the corresponding mechanisms, *whichever they are* (and they will not be ordinary perceptual processes), are not equally reliable. If an observational statement involved in the evaluation of an ordinary experiment is questioned, it is fair to invite the questioner to simply see for herself (perhaps after repeating the experiment). But the analogous reply in the case of a modal statement involved in the evaluation of a thought experiment — 'it's obvious; just think about it yourself' — clearly carries much less evidential and persuasive force." (pp. 72–73).

Here the corresponding mechanisms (see ft. 22) are the "psychological mechanisms linking the thought-experiment-as process to belief in pertinent premises [in particular C and $C \boxdot\!\!\rightarrow \neg W$]" (*ibid*). Translated into my account, the worry is that the OU is less supported than its counterpart observational statement in REs. Few paragraphs later, Häggqvist claims that this is the case for both TEs in science and philosophy, however without compelling arguments. I am more optimistic than Häggqvist though as to the intersubjective agreement for scientific TEs, at least an agreement on where to look for progress. That stems from my focusing on the theoretical statements, or lack of, that are used to describe, trace the execution of the experimental arrangement and interpret its result. To see this, consider the five strategies below. For the revealing part, ideally, when none of the first four strategies are available for the critic, an agreement should be reached, and sometimes is. We are then able to trust that the inconsistency revealed in the OU is serious for the theoretical statements involved. The first four strategies are meant to capture how the scientific community assess if the unfolding of the scenario (step 3) is *correctly done*. Put differently, the scientific community is assessing Häggqvist's claim 'it's obvious; just think about it yourself'.

7.2 Five strategies for the critic

The critic of a given TE disposes of at least the following five strategies.[23] The first four aim at challenging that an inconsistency is revealed, that is the OU follows from the unfolding. While the fifth acknowledges that an inconsistency is revealed, but aims at challenging its resolution by offering a different one.

7.2.1 The theoretical statements in 2(a) should be different

Bohr's reply to Einstein Solvay's photon-box TE is a canonical example. Briefly, we saw that Bohr proceeded by technically better-describing Einstein's weighting procedure and argued that the theories applicable are QM and GR, instead of QM and Gn. In Bohr's version *the OU no longer follows from the unfolding of the scenario*, thus blocking Einstein's attempt at revealing an inconsistency in QM. While in this case a better technical and theoretical description was deemed crucial, in some TEs the experimental details could be left under-described. Which brings me to the second closely related strategy.

7.2.2 Reject that the behaviour of some experimental arrangements could remain under-described

The history of Maxwell's demon is centred around better-describing the demon's mechanism, both technically and theoretically. These different TEs 'naturalize' the demonic process, either as a mechanical trapdoor (e.g., Smoluchowski 1912, Feynman et al. 1977) or as a computational measuring and erasing device (starting with Szilard 1929 until today). However, these naturalised TEs are *irrelevant for Maxwell's original demon* and *its conclusion.* As I have argued elsewhere (*cf.* El Skaf 2017 and El Skaf and Imbert 2013), these different TEs aimed at investigating if the second law should be *completely rejected* and if not why. While Maxwell only sought to argue that the second law is statistical in nature.

Bohr's reply could be analysed here if we consider that Einstein left his experimental arrangement under-described.

7.2.3 The theoretical statements are not correctly applied

This strategy is used mainly when the inconsistency is only *apparent*, and all is needed to block the inconsistency is to correctly apply the theoretical statements.

ETP (see Sect. 6) is an example. Not quite though, since Heisenberg principle needed to be extended to past events to block the inconsistency revealed by ETP, thus placing this TE in 7.2.1 above.

Langevin's twins TE (1911) and the Rocket and Thread TE (see Bokulich 2001) adopt this strategy. By correctly applying special relativity, both TEs resolve an

[23] These resemble the different schema from (β) (γ) and (δ) of how TEs could fail according to Häggqvist (pp. 66–67).

apparent inconsistency. For instance, Langevin argues that according to special relativity, *only* the travelling twin return to find that he is younger than his brother.

7.2.4 Reject the idealisations

The Aristotelian could reject the idealisation of the medium's resistance in Galileo's falling bodies TE, especially when the TE is taken out of context of Galileo's general argumentative strategy (*cf.* Palmieri (2005), El Skaf (2018) for a historical analysis of Galileo's falling bodies TE).

Norton (2018) provides a very interesting case study in analysing Szilard version of Maxwell's demon. He argues that the idealisation, which at first seemed harmless, turned out to be problematic when we look closer.

7.2.5 Propose a different resolution from the one proposed by the TEer

When a TE resist the criticisms following the first four strategies, the critic could instead focus on the resolution of the inconsistency proposed by the original author.

Einstein's proto-EPR photon-box is a canonical case: the TE conclusively showed that we should give up *either* locality/separability *or* completeness of QM (or a third alternative i.e., backward causation), all well-established physical principles. However, the TE does not provide any preference as to what to give up in the resolution step. Ironically, following Einstein criticism of QM's completeness, most scientist gave up locality and/or separability, without the question being settled to date.

The too many resolutions of Schrödinger's cat that are still to date. Like proto-EPR, Schrödinger's cat challenged QM's completeness (see Fine 1986, chapter 5). However, many resolutions kept completeness and tried to resolve the dead/living cat absurdity in different ways.

The several and sometimes incompatible research programs that followed the inconsistency revealed by black hole TEs. These programs span from QM to GR, from string theorist to relativists and promise to answer fundamental questions in physics such as the nature of gravity and its possible quantization.

Finally, energetists could (and some did) reject Maxwell's kinetic theory instead of limiting the second law of thermodynamics.

More generally, what I am defending in that the TEer offers a way out of the inconsistency only as a conjecture. This reflects the way they were, and still are, used by scientists. Indeed, I fail to see any given TE revealing an inconsistency and logically forcing a *unique* resolution; TEs, in and by themselves, lack any specified rule that forces scientists into a given resolution. This said, resolving inconsistencies, even as conjectures, is epistemically fertile. Different resolutions are proposed, which, when serious enough, act as starting point for new and sometimes incompatible research programs. These programs usually aim at further exploring and ideally confirming a resolution, thus enabling *scientific change and progress*. TEs thus have an important role to play in what Patton identified as the context of pursuit (see Sect. 7.1).

8 Conclusion

In this paper I proposed an account of many TEs that addresses their interrelated form, content and epistemic dimensions. The natural question that follows is do all TEs in physics (maybe beyond) share this analysis (at least all interesting ones)? For that we should inquire if the inconsistency revealing step 5 is present in all these TEs. Something I briefly mentioned in the introduction. For instance, some TEs aim at drawing implications from theoretical statements, at postulating a new general theoretical principle, or at arguing in favor of some other conclusions in step 6. To inquire if the inconsistency revealing step 5 is present even in these cases, a larger sample of case studies and a finer grained analysis of the different roles an inconsistency revealed could play is needed. This said, all TEs I have looked at until now (at least in physics) do include, in one way or another, this step 5. But arguing in favour of the generality of this account must be left to further research. A promising way would be to inquire further into the mutual relations and differences between TEs and different scientific tools, such as Scientific Models, Analogue Experiments and Computer Simulations. The presence of an inconsistency revealing step in TEs could possibly provide a demarcation criterion between TEs and other tools, in particular Scientific Models.

Acknowledgement I am indebted to Anouk Barberousse, John D. Norton, Mike T. Stuart, Margherita Archangeli, Sophie Roux, Patricia Palacios, Giovanni Valenete, Roman Frigg, Marcel Weber, Cyrille Imbert and anonymous referees for critical and helpful comments on earlier versions of this paper. Earlier versions of this paper were presented at several conferences, I would like to thank participants in Simulation and thought experiment (Geneva 2017), ECAP (Munich 2017), SoPhA (Louvain la Neuve 2018), LgBig (Geneva 2018), Fiction and Imagination conference (Torino 2019), Fiction, Understanding, and Thought Experiments workshop (Pairs 2019) Scientific and Epistemic Tools workshop (Salzburg 2020), for their helpful comments. This study was funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement No. 845008.

Funding Open access funding provided by Paris Lodron University of Salzburg..

References

Arcangeli, M. (2018). The hidden links between real, thought and numerical experiments. *Croatian Journal of Philosophy XVII*(52), 3–22.

Bishop, M. (1999). Why thought experiments are not arguments. *Philosophy of Science, 66,* 534–541.

Bohr, N. (1949). Discussions with einstein on epistemological problems in atomic physics. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-Scientist* (pp. 199–241). Cambridge: Cambridge University Press.

Bokulich, A. (2001). Rethinking thought experiments. *Perspectives on Science, 9*(3), 285–307.

Bokulich, A., & Frappier, M., et al. (2018). On the identity condition of thought experiments: thought experiments rethought. In M. T. Stuart (Ed.), *The Routledge Companion to Thought Experiments* (pp. 545–557). New York: Routledge.

Brendel, E. (2004). Intuition pumps and the proper use of thought experiments. *Dialectica, 58,* 88–108.

Brendel, E., et al. (2018). The argument view: are thought experiments mere picturesque arguments? In M. T. Stuart (Ed.), *The Routledge companion to thought experiments* (pp. 281–293). London and New York: Routledge.

Brown, J.R. (1991[2011]). *Laboratory of the mind: thought experiments in the natural sciences*, 2nd edition. London: Routledge.

Buzzoni, M. (2008). *Thought experiment in the natural sciences*. Würzburg: Königshausen and Neumann.

Canales, J. (2020). *Bedeviled: a shadow history of demons in science*. Princeton: Princeton University Press.

Cooper, R. (2005). Thought experiments. *Metaphilosophy, 36,* 328–347.

Einstein, A., Tolman, R. C., & Podolsky, B. (1931). Knowledge of past and future in quantum mechanics. *Physical Review, 47,* 777–780.

El Skaf, R. (2017). What notion of possibility should we use in assessing scientific thought experiments? *LatoSensu, 4*(1), 19–30.

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 XVII*(52), 37–58.

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

Feynman, R. P., Leighton, R.B., Sands, M. (1977). *The Feynman lectures on physics*, vol. I, original edition 1963. Boston: Addison-Wesley Publishing.

Fine, A. (1986). *The Shaky game: Einstein, realism, and the quantum theory*. Chicago: The University of Chicago Press.

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

Gooding, D. C. (1992). The cognitive turn, or, why do thought experiments work? In R. N. Giere (Ed.), *Cognitive models of science* (pp. 45–76). Minneapolis: University of Minnesota Press.

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

Howard, D. (1990). "Nicht Sein Kann was Nicht Sein Darf," or the prehistory of EPR, 1909–1935: Einstein's early worries about the quantum mechanics of composite systems. In Miller, A.I. (ed.) 1990*,* Sixty two years of uncertainty, NATO-ASI series B226. Plenum Press, pp. 61–111.

Krimsky, S. (1973). The use and misuse of critical Gedankenexperimente. *ZeitschriftfürallgemeineWissenschaftstheorie, 4,* 323–334.

Kuhn, T. (1964). A function for thought experiments, reprinted in *The essential tension*, T. Kuhn, Ed., (1977). Chicago: University of Chicago Press, 240–265.

Langevin, P. (1911). L'évolution de l'espaceet du temps. *Scientia, 10,* 31–54.

Maxwell, J. C. (1871). *Theory of heat*. London: Longmans, Green, and Co.

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

Meynell, L., et al. (2018). Images and imagination in thought experiments. In M. T. Stuart (Ed.), *The Routledge Companion to Thought Experiments* (pp. 498–511). London and New York: Routledge.

Nersessian, N. J. (1993). In the theoretician's laboratory: thought experimenting as mental modelling. In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *PSA 1992* (Vol. 2, pp. 291–301). Philosophy of Science Association: East Lansing.

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

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

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). Oxford: Blackwell.

Patton, L. (2012). Experiments and theory building. *Synthese, 184,* 235–246.

Popper, K. (1959). On the use and misuse of imaginary experiments, especially in Quantum Theory. In: *The logic of scientific discovery*. New York, Routledge (2005), 465–480.

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

Smoluchowski, M. V. (1912). Experimentellnachweis bare, der üblichenThermodynamikwidersprechende-Molekularphänomene. *PhysikalischeZeitschrift, 13,* 1069–1080.

Sorensen, R. A. (1992). *Thought experiments*. Oxford: Oxford University Press.
Stuart, M. T. (2016). Taming theory with thought experiments: understanding and scientific progress. *Studies in History and Philosophy of Science Part A, 58,* 24–33.
Stuart, M. T. (2019). Towards a dual process epistemology of imagination. *Synthese*. https://doi.org/10.1007/s11229-019-02116-w.
Stuart, M. T. (2020). The material theory of induction and the epistemology of thought experiments. *Studies in History and Philosophy of Science Part A, 83,* 17–27.
Szilard, L. (1929). Über, D. Entropieverminderung in einemthermodynamischen system beieingriffenintelligenterwesen. *ZeitschriftfürPhysik, 53,* 840–856.
Vickers, P. (2013). *Understanding inconsistent science*. Oxford: Oxford University Press.