Armchair physics and the method of cases

Abstract

Machery has recently argued that the unusual characteristics of philosophical thought experiments cause judgements about those thought experiments to be subject to the influence of extraneous factors such as demographic variables and order of presentation. Machery identifies the infrequent exposure to thought experiments as crucial for the unreliability of case judgments. In this paper we test this claim by gauging the reliability of judgments about thought experiments in physics. Since thought experiments in physics (Phy-TE) are structurally very similar to philosophical thought experiments (Phil--TE) and since Phy-TE have very similar characteristics as Phil-TE, judgements in Phy-TE should be unreliable too. Moreover, since physicists have comparatively little exposure to Phy-TE, their judgements should be unreliable too. However, we found that physicists are significantly better judges in Phy-TE than the folk. The difference cannot be explained by the influence of extraneous factors such as gender.

1 Introduction

Traditional philosophical theorizing is based substantially on judgements elicited in thought experiments. The content of these judgements is often used as (defeasible) evidence for or against philosophical views. This practice is also known as the 'method of cases'. Most prominently, Gettier cases are widely considered to have successfully refuted the theory of knowledge as justified true belief, Kripke's Gödel cases have played a key role in undermining naïve descriptivist theories of reference, and Putnam's Twin Earth has been used in support of semantic externalism, externalism about mental states, and essentialism about natural kinds. The nature of the judgements in thought experiments like these has been hotly debated. Although such judgements have often been described as "intuitions", both friends and critics of the method of cases have rejected any phenomenological characterization (Williamson 2007, Cappelen 2012, Machery 2017). Minimally, judgements in cases are case judgments (Machery 2017).

In the last decade, experimental philosophers have conducted experiments testing the case judgments of mostly non-professional philosophers, also known as the 'folk'. The findings indicate that the case judgements by the folk can vary with culture, gender, order of presentation and other seemingly irrelevant factors (Weinberg et al. 2001, Machery et al.

2004, Swain et al. 2008, Starmans and Friedman 2012, Buckwalter and Stich 2014). On the basis of such studies, experimental philosophers have questioned the unreflective use of the method of cases (Alexander and Weinberg 2007, Weinberg et al. 2010, Machery 2011, Alexander and Weinberg 2014, Machery 2017). Some have even concluded that case judgements are generally unreliable and that the method of cases should be given up entirely (Machery 2017).

In his recent book, Machery (2017) has argued that cases exhibit three types of "disturbing characteristics" which are likely to cause the unreliability of case judgments: cases are unusual, they pull apart properties that usually go together, and they contain superficial content irrelevant to the philosophical point in question. Importantly, not every case must exhibit *all* of these problematic features in order to cause unreliability, nor does any of the characteristics necessitate unreliability – they only make unreliability *more likely* (112).

Cases are unusual, according to Machery, "if and only if we encounter it infrequently or if we rarely read texts about it" (113). Trolley cases, for example, are unusual in that subjects are asked to decide over other lives; something that most subjects would not have had to consider at any time during their normal lives. Three features of a case's unusualness, Machery suggests, are likely to cause demographic and order of presentation effects. First, unusual situations may make it obscure "what facts hold in it in addition to the facts explicitly stated", for example in brain-swapping scenarios (114). Second, subjects may be "unable to grasp" the situations described in cases, so that "the thoughts people end up entertaining are not determined by the text they read" (114-115). Third, subjects cannot "rely on their memories of tried and true past judgments in everyday situations to get it right" or heuristics that have proven successful in everyday situations.

Philosophical cases, in Machery's view, also pull apart properties that go together in everyday life. For example, in the footbridge case, subjects are asked to engage in physical violence in order to do more good than harm, although violence usually causes more harm than good (116). In Gettier cases, subjects cannot rely on their usual strategies for identifying knowledge, as such cases pull apart properties epistemologists have identified as crucial for knowledge attributions (e.g. safety and adherence). Lastly,

philosophical cases are often described in "vivid terms", contain "many irrelevant narrative elements", and are presented in a "tendentious manner" (119).

Machery provides little evidence for his claim that the three disturbing characteristics he identifies *in fact* cause the unreliability of case judgements. For example, he provides no evidence that the pulling apart of properties in trolley *actually* causes subjects' susceptibility to order effects. Instead, Machery contends himself with having provided "*good explanations* of why the judgements examined by experimental philosophers are influenced by demographic and presentation variables" (112; added emphasis).

How could one go about testing Machery's claims about the characteristics of cases being disturbing? It is good practice in science to test hypotheses not only on the basis of the evidence which they were constructed for, but also with evidence that lies outside of their original domain of application. This avoids the suspicion of the hypothesis being gerrymandered to the evidence at hand and to actually get at something real {Worrall, 2002 #15}. Similarly, in philosophizing it is good practice to probe the claims of one's interlocutor by testing them against examples that satisfy the conditions set out by one's interlocutor but which were not considered and which might possess features which are capable of challenging one's interlocutor's claims. In this paper we apply this strategy to probe Machery's claims about the disturbing characteristics of cases. The example which we think satisfies at *least some* of these characteristics are *thought experiments in physics* (note that not more is required given that Machery doesn't think they all need to be present in a case for the case to cause unreliability).

It is important to note that Machery's claim about the disturbing characteristics of cases causing unreliable judgments is a general one: it applies to *anybody*. In fact, disturbing characteristics provide important reasons for Machery to reject the method of cases *tout court*. Yet, it is of course possible that the characteristics highlighted by Machery are more disturbing to some than they are for others. In particular, one might argue that philosophers should be less likely to be disturbed by the characteristics of cases than the folk, for the simple reason that philosophers deal with thought experiments on a daily basis. Thought experiments therefore should be much more usual to them than they are for the folk. Accordingly, their case judgements should be more reliable than the ones of the folk. The view that philosophers are better thought experimenters than the folk and

that their judgements are more reliable than the folk's is known as the *expertise defense* (Hales 2006, Ludwig 2007, Horvath 2010, Devitt 2011, Williamson 2011). Interestingly, Machery himself admits that the "proper domain" of reliable judgments "varies with the expertise of the person judging" (112). However he provides strong arguments against the idea that philosophers have any special expertise in thought experimentation.

Thought experiments in physics are of course not used as widely physicists as philosophical thought experiments are used by philosophers. Thought experiments in physics should thus be similarly disturbing to physicists as they are to laypeople. Both the laypeople's and the physicists' case judgements should accordingly be unreliable. Thought experiments in physics therefore provide a nice test case for Machery's account: should physicists' case judgments turn out to be reliable and the folk's unreliable, the allegedly disturbing characteristics of cases cannot be the causes of unreliability (as they are disturbing for both groups). Contrary Machery, the characteristics of thought experiments could thus not be invoked to argue for the unreliability of case judgements.

Our discussion is structured as follows. In Section 2 we establish that thought experiments in physics are a relevant class of examples for testing Machery's claim about the adverse effects of disturbing characteristics on the reliability of case judgments. Crucially, we argue that if thought experiments in philosophy really exhibit disturbing characteristics (as Machery claims), then thought experiments in physics do so too. Case judgments in thought experiments in physics should accordingly be unreliable as well. In Section 3 we present the methods used in our test of the hypothesis that physicists make more reliable case judgments than non-physicists. In Section 4, we present the results of our study. In Section 5 we draw our conclusions.

2 Armchair physics

Although clearly not as central as in philosophy, thought experiments have played an important role in the history of physics, particularly in periods in which new theories were proposed and old ones rejected (Kuhn 1977).[1] For example, in the mid-16th century Galileo famously showed that there was an inherent contradiction in Aristotelian physics by

[1] Interestingly, thought experiments seem to be rare in other sciences such as biology and chemistry. But see Stuart et al. (2017) for some interesting examples from outside the realm of physics.

means of a thought experiment of two connected falling objects of different weights (Gendler 1998). In the late 1920s, "Schrödinger's cat" was used by Schrödinger and Einstein to argue against the plausibility of the Copenhagen interpretation of the new quantum mechanics.

In this section we will first establish that the structure of thought experiments in physics and the way in which case judgments figure in them are relevantly similar to thought experiments and case judgments in philosophy (2.1). In particular, we will argue that thought experiments in physics also exhibit some of Machery's 'disturbing characteristics' (2.2). We will then make a case for case judgments in the practice of physics playing a similarly (albeit much more limited) evidential role as case judgments in the practice of philosophy (2.3). We also argue that the epistemology of those case judgements in physics should be comparable to the one in philosophy (2.4). At the end of this section, we discuss potential differences between case judgments in philosophy and physics (2.5).

2.1 Thought experiments in physics and in philosophy compared

There is a plethora of views regarding the nature of thought experiments in physics (Phy-TE). They range from constructivist views according to which thought experiments are capable of pointing us to weaknesses of current theories fail and how to fix them (Kuhn 1977, Gendler 1998, Camilleri 2014), views which construe thought experimentation as a form of mental modelling (Nersessian 1992), to Platonist views which conceive of some thought experiments as windows to laws of Nature, understood as relations of necessitation (Brown 1991). Minimally, thought experiments in physics are arguments, or at least can reconstructed as such (Norton 2004). Indeed, many important thought experiments in both physics *and* philosophy have the following argument form (Häggqvist 2009):

1. *The imaginary scenario*: It is possible that A(ntecedent) is the case.
2. *The case judgement (as counterfactual conditional)*: If it were the case that A, C(onsequent) would be false.
3. *The theoretical conditional:* If theory T is true, then, if it were the case that A, C should be true.
4. *Conclusion (by modus tollens)*: Therefore, T is false.

That is, an important class of thought experiments consists of the consideration of an imaginary / counterfactual scenario, in which A is the case and where it is judged (and counterfactually) that C is not the case. Since T entails C when A is the case, T is falsified by the thought experiment in question. The evidence for the falsehood of T is thus provided by the counterfactual judgement that C. Consider for example the following reconstruction of Schrödinger's cat:[2]

S1. It is possible that a cat can be situated in a way described by the thought experiment.
S2. If the cat were situated in this way, it would have to be either alive or dead.
S3. If the Copenhagen interpretation is true, then, if a cat were situated this way, it would be both dead and alive.
S4. Therefore, the Copenhagen interpretation is false.

This seems very similar to the form of many Phil-TE, for example, Gettier cases:

G1. It is possible for somebody (S) to be in a Gettier situation with regard to p.
G2. If S were in a Gettier situation with regard to p, S would justifiably believe the true proposition p without knowing that p.
G3. If knowledge is justified true belief, then if S were in a Gettier situation with regard to p, S would know that p.
G4. Therefore, knowledge is not justified true belief.

In Phil-TE such as the Gettier cases, the relevant case judgement (here: G2) figures as evidence in the attempted refutation of the theory in question (here the theory that knowledge is justified true belief). Likewise, in Phy-TE such as Schroedinger's cat, there is a judgement (here: S2) which seems to perform an evidential function: it is by virtue of judgements like S2 that the theory in question is being rejected (here: the Copenhagen interpretation). We should hasten to add that this is no curiosity of Schroedinger's cat. Consider also a reconstruction of the clock-in-the-box thought experiment, which played an important role in the debate between Einstein and Bohr on the plausibility of the uncertainty principle (from Häggqvist (2009)):

[2] We note that we here represent this thought experiment with the conclusion intended by its originator. Another representation would deny the conclusion S4 and maintain the truth of the Copenhagen interpretation by rejecting the case judgement S2. Yet another representation would deny the theoretical conditional S3, namely that a state of superposition of macro-objects such as a cat is entailed by the Copenhagen interpretation in the scenario in question.

E1. It is possible (in principle) that a single photon exit a box equipped with an arbitrarily exact timer and an arbitrarily sensitive spring-balance.

E2. If the uncertainty principle holds, then if a single photon exited a box equipped with an arbitrarily exact timer and the box were then weighed, the time and energy of its passage would not be simultaneously measurable to any degree of accuracy violating the inequality $\Delta E \times \Delta t > h$ (where h is Planck's constant / 2π).

E3. But if a single photon exited a box equipped with an arbitrarily exact timer and the box were then weighed, the time and energy of its passage would be simultaneously measurable to any degree of accuracy.

E4. Hence, the uncertainty principle doesn't hold.

Again, the thought experiment closely follows the general structure laid out above. E1 describes the imaginary scenario of the thought experiment, E2 is the theoretical conditional that sets out the uncertainty principle as the target of the thought experiment, E3 is the case judgement (in the form of a counterfactual conditional), and E4 concludes that the uncertainty principle is false *on the basis of the judgement* (E3) about the pondered scenario (E1). Thus, crucially, the case judgement performs an evidential function in the thought experiment.

More specifically and more correctly, it should be emphasized, the evidence in thought experiments is not so much provided by the case judgement itself, but rather by the *content* of the case judgement (Williamson 2007, Machery 2017). To see this, consider for example the argument in which the *judgement* that "Goedel" refers to Goedel and not Schmidt is the premise for the conclusion that "Goedel" refers to Goedel and not Schmidt (which would in turn figure in an argument against descriptivism). In such an argument, we would have to assert the conclusion *before* we can assert the premise, since we cannot make a judgement that X without X already being available. Likewise, if the judgment that X could be evidence for one's belief that X, then it would be possible to bootstrap oneself to certainty that X: once could incrementally increase one's confidence that X simply by repeatedly *judging* that X. Obviously these lessons hold for both Phil-TE *and* Phy-TE. For simplicity's sake, however, we shall continue to speak of the evidential function of judgements.

We should note that in the particular thought experiment just considered, it would eventually turn out that the evidential function assigned to E3 by Einstein was mistaken:

after some hard thinking (see Section 2.3), Bohr realized that E3 leaves out of consideration relativistic effects which result from the weighing of the box. The consequent of the counterfactual conditional is therefore not correct (as suggested in E3). This shows us that case judgements in Phy-TE – just as case judgements in Phil-TE – are of course fallible.

2.2 Disturbing characteristics in Phy-TE

It is quite apparent that many Phy-TE exhibit the three disturbing characteristics of thought experiments identified by Machery, namely unusualness, the pulling apart of properties that usually go together, and superficial content irrelevant to the point in question. In the two examples already mentioned, both Schroedinger's cat and Einstein's clock-in-the-box are unusual in that these are not the kinds of situations physicists usually ponder. Both thought experiments describe imaginary scenarios which are practically impossible: it's impossible to build an apparatus that would measure a *single* photon and it's impossible to determine the state of the cat in the box without employing some sort of measurement. Likewise, in Newton's cannon, it is practically impossible (for the conceivable future) to elevate a cannon high enough and to generate the energy that would be required to shoot a cannon ball into an orbit around the earth. Also, in Einstein's famous elevator TE, it seems impossible (by our means) to pull an elevator with a person inside with the required speed through space that has no nearby gravitational fields (see Appendix 2 for details). Although there are of course always technological limit to the experiments that physicists can perform, the scenarios described by thought experiments in physics are often so outlandish that there is not even a remote possibility that there ever will be experiments that will probe such scenarios. In that sense, Phy-TE differ very much from hypothetical experiments that physicists might consider even before the required technological means are available to them. Phy-TE are thus surely unusual.

Just like Phil-TE, many Phy-TE pull apart properties that usually go together. For example, quantum properties are usually associated with micro-objects like electrons. Schroedinger's cat asks us to consider a scenario in which quantum properties are assigned to the macro-object of a cat. Usually, projectiles never have the speed they have that would be required to shoot an object into an orbit around the earth.

Lastly, many Phy-TE, just like Phil-TE, offer lots of 'irrelevant narrative elements' in what Machery calls 'superficial content' of a thought experiment. For example, it is

entirely irrelevant whether the cat in the box dies because a Geiger counter or some other device triggers the release of a toxin, or whether the toxin is contained in a flask that breaks. Likewise, it is quite irrelevant to Newton's cannon whether the device used to shoot the object into an orbit around the earth is an (unrealistically powerful) cannon or an (unrealistically powerful) rocket launcher.

Thus, many Phy-TE seem to exhibit precisely those properties which Machery has identified as problematic in Phil-TE. If Machery is right, those properties should cause the *general* unreliability of case judgments also in Phy-TE. What's more, given that Phy-TE even in the practice of physics seem to be rather rare, they should be unusual *even to physicists* (in sharp contrast to Phil-TE not being unusual to philosophers at all). Whether Phy-TE cause the unreliability of judgments for *any* subject, we set out to test in this paper. But first, we want to consider in more depth the similarities of case judgments in both Phy-TE and Phil-TE.

2.3 Case judgements as evidence in the practice of physics

In the history of the discipline of philosophy judgements elicited in Phil-TE have been instrumental in rejecting previously widely held philosophical views. For example, the relevant judgements in the Gettier cases have been instrumental in undermining the JTB theory of knowledge, Kripke's Goedel cases have helped undermine descriptivism, etc. Although it has of course been contested by experimental philosophers that judgements elicited by thought experiments *ought to* be used as evidence, it is denied by few that they have *in fact* played an evidential role in philosophical practice. That is, it is denied by few that case judgements have historically constituted important reasons for rejecting certain philosophical accounts and supporting others (see e.g. Machery (2011, 2017)).

In what sense—if at all—have judgements in Phy-TE played an evidential role in the practice of physics? Clearly, physicists have better evidence sources at their disposal than judgments in TEs, on which they can rely when assessing theories. Still, there are historical episodes in which thought experiments and the judgements elicited by them *have* had an evidential function in scientific debates. For example, in their famous debate about the foundations of quantum mechanics, Einstein advanced the aforementioned "clock in the box" thought experiment in which he challenged Bohr on the plausibility of Heisenberg's uncertainty relation by means of the so-called "clock in the box" thought

experiment. Bohr's first reaction and his eventual victory in the argument with Einstein has been reported in the following way:

> It was quite a shock for Bohr ... he did not see the solution at once. During the whole evening he was extremely unhappy, going from one to the other and trying to persuade them that it couldn't be true, that it would be the end of physics if Einstein were right; but he couldn't produce any refutation. I shall never forget the vision of the two antagonists leaving the club: Einstein a tall majestic figure, walking quietly, with a somewhat ironical smile, and Bohr trotting near him, very excited. ... The next morning came Bohr's triumph. (Rosenfeld, cited in Pais 1982, 446-7)[3]

Bohr accepted that Einstein's thought experiment elicited the judgement also in him and that the uncertainty relation would be violated in such a scenario. His reported reaction illustrates how seriously he took this thought experiment as *evidence* against the uncertainty principle. Luckily for him, he found a way to avoid this threat (cf. Bishop 1999). As mentioned, thought experiments have played comparable roles particularly in the early stages of the development of theories, in which the conceptual coherences of new theories are explored and in which evidence from real experiments is sparse or hard to produce (cf. Kuhn 1977).

2.4 The epistemology of case judgements in Phy-TE

If case judgements in thought experiments can be evidential in the ways described, what are the conditions under which they are true or false (and therefore evidential and non-evidential, respectively) and how are such truth values determined? In many cases in Phy-TE, the truth of the counterfactuals that case judgements arguably consist of cannot be determined on an experimental basis – again, the lack of the appropriate experimental resources is often a reason for resorting to thought experiments in the first place! Moreover, many thought experiments in physics cannot even possibly be carried out in our world (as already noted in Section 2.1). But if there is no experimental evidence to determine the result of (some) Phy-TE and (some) Phy-TE are impossible to conduct in this world, then, again, how do physicists evaluate the truth content of judgements about Phy-TEs? In a similar way as philosophers evaluate the truth content of judgements in Phil-TE, it seems. As mentioned above, standard accounts reconstruct thought experiments as counterfactual arguments and the ability to 'conduct' thought experiments

[3] See also Bishop (1999) for a philosophical discussion.

as the ability to evaluate counterfactual conditionals (Norton 1991, Williamson 2007, Häggqvist 2009). Roughly, when one engages in counterfactual reasoning, one 'enriches' the antecedent via mental simulation and with the help of one's background knowledge. If this enrichment results in the truth of the consequent of the counterfactual, the whole counterfactual conditional is acceptable. This seems to be a plausible model not only in Phil-TE, but also in Phy-TE. Also in Phy-TE, on has to resort to one's imagination and our background knowledge in order to produce the relevant scenarios before "one's eyes". Also in Phy-TE, one needs to judge whether the enrichment of the antecedent of the relevant counterfactual leads to the truth of the consequent. In accordance with the Lewis-Stalnaker semantics of counterfactuals, physically possible worlds should be considered closer to the actual world than physically impossible but metaphysically possible worlds. Whereas the latter violate the laws of nature, the former don't. The metric and the procedure for evaluating the closeness of any of such worlds, though, should be the same.[4] On the side of Phil-TE, one should note that even when we are interested in metaphysically possible worlds, the evaluation of the relevant counterfactual is arguably *a posteriori* (rather than a priori), as we will still have to consult our background knowledge about the actual world in order to be able to judge whether the worlds are sufficiently similar (Williamson 2007). The epistemology of case judgements in both Phy-TE and Phil-TE thus looks very similar.

2.5 (Potential) differences between judgements in Phy-TE and Phil-TE

Let us now consider senses in which Phy-TE and Phil-TE might be disanalogous. We will take our lead from Nado (2014a, 2015), who, in a pair of interesting papers on the expertise defense, has argued recently that philosophical judgments are different from scientific judgements in at least two senses. Nado does not consider case judgements in Phy-TE (as we do), but rather just judgements in science. Nevertheless, her critique may be thought to extrapolate.

First, Nado claims, judgements in science and case judgements in philosophy different because in science, practitioners have *explicit* access to the well-established

[4] According to the standard account, one evaluates the truth of a counterfactual by evaluating by means of a similarity metric whether the consequent is true in the possible world closest to the actual world in which the antecedent is true.

principles underlying their judgements, whereas philosophers don't. For example, when we judge the shape of trajectory a cannon ball will take when shot into the distance, the physicist can of course justify her judgement in a way that philosophers cannot, namely by appealing to the well-established principles of Newtonian mechanics. In contrast, philosophers cannot appeal to such principles when evaluating thought experiments. Second, Nado claims, whereas judgements in Phil-TE serve as evidence for or against philosophical theories, this is not the case in physics where judgements (about the trajectory of a projectile will take) are *supported* by well-founded theory. The evidential relation thus seems to go in the opposite direction of the one in Phil-TE.

Let us start with Nado's second concern. As we mentioned before, thought experiments in physics (again, Nado does not consider those explicitly) are normally advanced in periods in which the theoretical foundations of science are in flux (Kuhn 1977). In such periods, scientists cannot appeal to the truth of the relevant theories and principles in order to justify their judgements. On the contrary, scientists use thought experiments in these periods in order to lend plausibility to the theories they want to defend (or in order to challenge the theories they want to overcome). Galileo, for example, sought to argue for the heliocentric system by appealing to a thought experiment in which a cannon ball was dropped from the top of the mast of a moving ship. The judgment the thought experiment was supposed to elicit was that the ball should drop to the bottom of the mast and not somewhere behind the mast (in the direction opposite the ship's direction of motion), because the ball, by virtue of being part of the (moving) inertial system of the ship, would not only have a vertical motion but also a horizontal one. Galileo used this thought experiment to argue against those who believed that a moving earth would imply falling objects to fly in the direction opposite the earth's direction of motion.

In sum, Nado's second objection to the analogy underlying the expertise defense does not apply to our version of the analogy, because just like in case judgements in Phil-TEs, judgements in Phy-TEs serve as evidence for theories, and not the other way around. Of course, physicists were later able to gather kinds of evidence other than judgements in Phy-TE to support their theories. But at the time at which judgments in Phy-TE were used, and at which they had their greatest evidential force, this other evidence wasn't available.

We do not think that Nado's first concern really applies to case judgements in Phy-TE either. Are the principles that underlie those judgements always transparent and accessible to those making the judgments? Did Aristotelians, for example, really have explicit access to the principle of Galilean relativity when they were asked to judge where the ball dropped from the top of the ship mast would land? Was the principle transparent to them? That seems highly unlikely, as not even Galileo himself had no way of establishing this principle other than by way of examples like these (which is presumably one of the reasons why he used them in the first place). It might even be questioned whether case judgements in Phil-TE are really as "frustratingly opaque" as they are portrayed by Nado. In the Gettier cases, for example, we have all developed an appreciation that one cannot know by luck and that it is this 'principle' which underlies our judgement that "Smith doesn't know".

We conclude that there are a number of important and interesting similarities between Phil-TE and Phy-TE. This is of course not to say that there are no differences whatsoever between the two. For example, it seems undeniable that theoretical knowledge about the world plays a bigger role in Phy-TE than it does in Phil-TE. In fact, this knowledge can often conflict with common sense in a way that philosophical views often don't (McCloskey 1983b, a, McCloskey and Kohl 1983). In contrast, philosophical theories are often built in such a way so as to retain consistency with common sense (although some philosophers have argued recently that we should give up on this goal; cf. Hawthorne and Fairchild (forthcoming)). We will further explore this difference in the conclusion of this paper.

3 Method

As we mentioned in Section 2.2, given the low frequency of Phy-TE in the practice of physics, the characteristics of Phy-TE should be disturbing even for physicists, if Machery's account is correct. Accordingly, not only the case judgements of laypeople should be unreliable, but also the case judgments of physicists. This we sought to test.

In order to test the reliability of case judgements, we relied on an idea used by Horvath and Wiegmann (2015) in the context of testing case judgements in Phil-TE: we used the textbook consensus about case judgments in Phy-TE as a standard for gauging

the reliability of the case judgements. This standard is readily available (Brown and Fehige 2011).

3.1 Materials

We designed a set of six tasks. Each task consisted of a description of a Phy-TE, a figure representing the Phy-TE, a comprehension question about the text, and a statement describing either the standard judgement about the thought experiment, or its negation. We chose six classical Phy-TE for our tasks, which are well known and discussed in the philosophy of science literature (Brown and Fehige 2011): Stevin's chain, Schroedinger's cat, Galileo's tower, Galileo's ship, and Newton's cannon. Participants were asked to carefully consider the thought experiment and answer two questions. The first question was a relatively simple comprehension question that asked participants to finish a statement about a relevant element of the scenario and was designed to probe whether they have understood the text. Only participants who answered the comprehension question correctly were included in our analysis. The comprehension questions were designed in such a way that they would not guide the participants towards the correct answers to the second question (see Appendix 2).

The second question asked participants to what extent they agreed or disagreed with a statement expressing a judgment in the hypothetical scenario. In half of the tasks participants were asked to evaluate statements expressing a correct judgement, and in the other half of the tasks participants were asked to evaluate the statement expressing the negation of a correct judgement. The participants were asked to answer by indicating their level of agreement/disagreement on a five-point Likert scale: 1 = Strongly disagree; 2 = Somewhat disagree; 3 = Neither agree nor disagree; 4 = Somewhat agree; 5 = Strongly agree. We measured those scores against the standard judgements reported in the literature on Phy-TE (Brown and Fehige 2011). For example, in the hypothetical scenario based on the Schrodinger's cat thought experiment, subjects were asked to indicate to what extent they agreed/disagreed with the statement "Before the box is opened the dog is either dead or alive".[5] The six tasks were presented randomly and were followed by a short questionnaire gathering (see next section). For details concerning tasks and their

[5] In our experiment, we assumed that physicists would reject this judgement. See footnote 2 and Section 4 and Appendix 2 for further details.

design see Appendix 2. In order to test the idea that physicists are more reliable than the folk in judging thought experiments, we tested whether physicists are more likely to judge our six tasks correctly than the folk (as measured by the textbook standard). Our hypothesis was therefore the following:

H: Physicists are more likely to make correct judgements in Phy-TE than non-physicists.

If H is correct, Machery's claim that thought experiments exhibit features that are disturbing to subjects and which cause the unreliability of case judgements would be challenged, because the relevant features of Phy-TE should be disturbing also for physicists (who are rarely exposed to such features).

One may think that H is prima facie plausible, given that physicists should have knowledge in physics which the folk don't possess and which is required for correctly judging Phy-TE. However, it is not obvious that the amount of physics required for correctly judging Phy-TE would not be available to non-physicists. For example, Galileo's tower only requires a grasp on the (basic) logic required to deduce a *reductio*. Similarly, Phy-TE in general do not impose very high computational demands on subjects: in contrast to many other physics tasks, one need not solve any mathematical equations in order to generate the correct judgement. Lastly, it is interesting to note that even in basic tasks in classical mechanics, subjects trained in physics can make systematic mistakes (McCloskey et al. 1980, Oberle et al. 2005).

3.2 Participants

We tested two groups of subjects: those with and those without (or only minimal) university-level education in physics. In order to ensure that we would get experts in physics, we made it a requirement that our subjects had a PhD degree or were studying towards one. The participants in the physics group (n = 57) were thus recruited via mailing lists for PhD students, postdocs and faculty members in various physics departments in <blinded for peer review>. The mean age of participants in the physics group was 33 (SD = 10.2) of which 11 (19%) were female.[6] Out of all 57 participants in this group, 32 (56.1%) reported having a PhD and 25 (43.9%) being enrolled in a PhD programme in physics (see Appendix 1 for specialisations).

[6] This reflects the unfortunate underrepresentation of women in physics (Sax et al. 2016)

The participants in the control group (n = 57) were recruited via mailing lists for PhD students, postdocs and faculty members in various departments in science, social science, and humanities and excluding physics departments in <blinded for peer review> (see Appendix 1 for details). The biggest subgroup were political scientists (n=19). The mean age of participants in the control group was 34 (SD = 6.72).The control group consisted of 57 subjects of which 27 (47.3%) were female, 33 (57.9%) reported holding a PhD and 24 (42.1%) being enrolled in a PhD programme. The details regarding their education and areas of study are presented in Appendix 1. A background section located at the end of the questionnaire was used to collect information on participants' education, age, gender, level of English, and area of specialisation.[7]

4 Results and discussion

In order to test our hypothesis that physics experts are more likely to judge Phy-TEs correctly than non-physicists, we determined the average number of correctly answered Phy-TE for both the physics experts and our control group. We distinguished between a STRICT and a LAX condition, whereby STRICT stands for strong agreement or strong disagreement with the presented judgement (which could be either correct or false) and LAX includes both the "strongly agree/disagree" and the "somewhat agree/disagree" responses. We then determined whether the difference in the means of the number of correct judgements by physics experts and the non-physicists was highly significant in these two conditions with a T-test. For both conditions, our hypothesis was confirmed: physicists on average judged correctly about one task more than the non-physicists in the LAX condition and about one-and-a-half task more in the STRICT condition (see Table 1).

	Physicists, mean	SD	Non-Physicists, mean	SD	*p*	t
LAX	4.30	1.12	3.37	1.24	.000	-4.22
STRICT	3.44	1.23	2.02	1.40	.000	-5.89

[7] In total, 163 participants responded to our call and submitted their responses via the *Qualtrics* platform. Each participant had the option to leave their email address on an external Google Forms website in order to enter a lottery for 5 amazon.com vouchers of each $25 or to receive information about the results of the study. We excluded partly incomplete questionnaires (n = 29), subjects who did not satisfy our minimal criteria for education, i.e. not being enrolled in a PhD programme (n = 13), subjects who did not have at least an intermediate level of English (n=1), and subjects who did not answer the comprehension questions (n=6).

Table 1: Average number of correctly answered tasks by physicists and non-physicists out of a total of six tasks.

We conducted negative binomial regression analysis to test for the influence of gender, age, response duration, exposure of controls to physics at university, and level of English (see Appendix 4). None of these factors was significant. That means that the difference we found between the performance of the physicists and the non-physicists cannot be accounted for by at least one of the extraneous factors which have been reported to influence case judgements in the experimental philosophy literature, namely gender (see Introduction). Physicists' higher degree of reliability is probably best explained by their better knowledge of the physical principles at play in our six tasks.

Apart from the overall result, we also analyzed the correct judgements for each of our six thought experiments in both LAX and STRICT with a chi-squared test (Table 8 in Appendix 3). Table 2 lists the outcomes of these tests together with the percentages of correct judgements for each thought experiment. There is a big variance across the tasks: some tasks were answered correctly by almost all physicists (in particular Galileo's tower), whereas other tasks had many physicists answer incorrectly. Schroedinger's cat stands out from the other tasks in that most subjects in both groups answered contrary to our expectation. We will discuss this case in more detail in a moment.

	% correct in LAX		*p*	% correct in STRICT		*p*
	Physicists	Non-physicists		Physicists	Non-physicists	
Stevin's chain	68.4	59.6	ns	56.1	38.6	ns
Einstein's elevator	59.6	54.4	ns	50.9	31.6	*
Schroedinger's cat	26.3	10.3	*	14.0	7.0	ns
Newton's cannon	82.3	68.4	ns	57.9	24.6	***
Galileo's ship	94.7	66.6	***	70.2	42.1	**
Galileo's tower	98.2	77.2	**	94.7	57.9	***

Table 2: percentages of correct responses in both the LAX and the STRICT condition for both physicists and non-physicists. Statistically significant differences from our χ^2 test at the 95% level are marked with a star (*), at the 99% level with a double-star (**) and at the 99.9% level with a triple-star (***). See Table 8 in Appendix 3 for details regarding the tests.

An obvious question one must ask is: can the variance of our results be explained by some thought experiments being more disturbing than others? This question might not have a simple answer. Newton's cannon, for example, seems to be a straightforward application of the principles of classical mechanics. Yet, physicists are not perfect in this task, and many do not strongly agree with the correct judgment. The reason why they don't may indeed have something to do with the disturbing characteristics of the case (see Section 2.2). Stevin's chain, on the other hand, seems to be a perfectly normal scenario: disturbing characteristics are pretty much absent. Still, a significant amount of physicists did not answer the task correctly in either the LAX or the STRICT condition. So at the very least, the answer to the question of the influence of disturbing characteristics does not seem to be a consistent one. In fact, the subpar performance of physicists in some tasks may have to do with something else entirely. Take for example Einstein's elevator. Only about half of our physicists answered correctly in the STRICT condition. This may have to do with the scenario requiring a particular acceleration of the elevator for the person in the elevator not to be able to distinguish between the pull of the elevator and the gravitation exerted by earth. In other words, some of our Phy-TE may be too underspecified for physicists to make confident case judgements. At any rate, these are speculations. Our test did not have the right 'resolution' to shed light on the reasons of our subjects' performance in each single task. Instead, our test was designed to find out about our subjects' performance across *several* thought experiments. We therefore re-emphasize that our test result confirmed our hypothesis that physics experts were more reliable than the non-physicists. This, again, is at odds with the idea that the characteristics of thought experiments should be disturbing to both groups, given the comparatively infrequent exposure of physicists to thought experiments.

Let us return to Schroedinger's cat. Again, this task sticks out from the other tasks in that most physicists came out as answering this task incorrectly in both of our conditions. In order to try to elucidate this result, we will provide some further background on the original motivation for this thought experiment, which shaped our expectation of how physicists would respond to this thought experiment.

Schrödinger used his thought experiment to challenge Bohr and Heisenberg's Copenhagen interpretation of quantum mechanics. Schroedinger's reasoning was that if

the Copenhagen interpretation were correct, then the cat in the box (in our case: a dog) should be in a state of superposition before the opening of the box (the "measurement") causes a collapse of the wavefunction. However, since we would under normal circumstances judge that the cat/dog does have a definite state before we open the box, the Copenhagen interpretation must be false. Since the Copenhagen interpretation is indeed by far the most accepted interpretation amongst physicists[8], we expected physicists to 'bite the bullet' and accept that the cat/dog actually is in a state of superposition.[9] It turned out, however, that most physicists did not judge this way (even though they did judge this way significantly more often than the folk). They therefore presumably rejected that the Copenhagen interpretation (contrary to what Schroedinger thought) *entails* the judgement that the cat/dog must be in a state of superposition. In other words, physicists seemed to reject what we called the "theoretical conditional" above in our reconstruction of the Schroedinger cat thought experiment (see p. 5). From our pilots, we gathered that physicists did this on the basis of making a distinction between micro-objects such as electrons and macro-objects, such as cats and dogs. We take it that such a distinction can be made within the Copenhagen interpretation on the basis of Bohr's correspondence principle, which says (roughly) that the predictions of quantum mechanics approximate those of classical mechanics when it comes to classical objects (such as cats and dogs).

We should note that even when reversing the scales for our Schroedinger cat task so that (strong) agreement with the presented judgement is assumed to be correct (instead of incorrect as in our original analysis), our hypothesis would remain confirmed (see Table 3).

	Physicists, mean	SD	Non-Physicists, mean	SD	*p*	t
LAX	4.60	.979	4	1.363	.003	-2.68
STRICT	3.86	1.187	2.68	1.549	.000	-4.55

Table 3: Average of correctly answered questions out of a total of six tasks, with the scale for the Schroedinger cat reversed ($p < .001$ for STRICT and $p < .005$ for LAX).

[8] For a recent survey see Sivasundaram (2016).

[9] We were influenced by a standard textbook in the philosophy of physics (Sklar 1992, 184).

At the same time, it should be pointed out that this scale reversal would result in the non-physicists answering correctly more often than the physicists (see Table 4). This would make the Schroedinger cat the only of our six tasks in which this would be the case.

	% correct in LAX		% correct in STRICT	
	Physicists	Non-physicists	Physicists	Non-physicists
Schroedinger's cat (reversed scales)	66.7	82.5	56.1	73.7

Table 4: Percentages of correct answers for Schroedinger's cat with reversed scales in both LAX and STRICT. LAX: X^2 (1, 114) =3.746 (p=.053); STRICT: X^2 (1, 114) =3.851 (p=.050).

5 Conclusion

This paper tested Machery's recent claim that philosophical thought experiments exhibit disturbing characteristics that cause case judgements to be unreliable. We conducted this test on the basis of thought experiments in physics, which we argued share crucial features with thought experiments in philosophy. We also tested the reliability of both physicists' and non-physicists' case judgements. Our result of physicists' case judgements being more reliable than the folks' challenges Machery's view that the characteristics of thought experiments should be disturbing for anybody. Insofar case judgments in philosophy are unreliable (as claimed by Machery and others), they are probably not unreliable by virtue of the characteristics of thought experiments.

References

Alexander, J. and J.M. Weinberg. 2007. Analytic epistemology and experimental philosophy. *Philosophy Compass*, **2** (1): 56-80.

———. 2014. The "unreliability" of epistemic intuitions. In *Current Controversies in Experimental Philosophy*, Edouard Machery and Elizabeth O'Neill (eds.), New York: Routledge, 128-145.

Bishop, M.A. 1999. Why thought experiments are not arguments. *Philosophy of Science*: 534-541.

Brown, J.R. 1991. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. New York: Routledge.

Brown, J.R. and Y.J.H. Fehige. 2011. Thought experiments. *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, http://plato.stanford.edu/archives/fall2011/entries/thought-experiment/.

Buckwalter, W. and S. Stich. 2014. Gender and Philosophical Intuition. In *Experimental Philosophy*, Joshua Knobe and Shaun Nichols (eds.), Oxford: Oxford University Press.

Camilleri, K. 2014. Toward a constructivist epistemology of thought experiments in science. *Synthese*, **191** (8): 1697-1716.

Cappelen, H. 2012. *Philosophy without intuitions*. Oxford: Oxford University Press.

Devitt, M. 2011. Experimental semantics. *Philosophy and Phenomenological Research*, **82** (2): 418-435.

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

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

Hales, S.D. 2006. *Relativism and the Foundations of Philosophy*. Cambridge, MA: MIT Press.

Hawthorne, J. and M. Fairchild. forthcoming. Against Conservatism in Metaphysics. *Philosophy*.

Horvath, J. 2010. How (not) to react to experimental philosophy. *Philosophical Psychology*, **23** (4): 447-480.

Horvath, J. and A. Wiegmann. 2015. Intuitive expertise and intuitions about knowledge. *Philosophical Studies*: 1-26.

Kuhn, T.S. 1977. A function for thought experiments. In *The Essential Tension*, Thomas S. Kuhn (ed.), Chicago: University of Chicago Press, 240–265.

Ludwig, K. 2007. The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, **31** (1): 128-159.

Machery, E. 2011. Thought experiments and philosophical knowledge. *Metaphilosophy*, **42** (3): 191-214.

———. 2012. Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, **27** (1): 37-54.

———. 2017. *Philosophy within its proper bounds*: Oxford University Press.

Machery, E., R. Mallon, S. Nichols, and S.P. Stich. 2004. Semantics, cross-cultural style. *Cognition*, **92** (3): B1-B12.

McCloskey, M. 1983a. Intuitive Physics. *Scientific American*, **248** (4): 122-131.

———. 1983b. Naive theories of motion. *Mental models*: 299-324.

McCloskey, M., A. Caramazza, and B. Green. 1980. Curvilinear Motion in the Absence of External Forces: Naive Beliefs about the Motion of Objects. *Science*, **210** (4474): 1139-1141.

McCloskey, M. and D. Kohl. 1983. Naive physics: the curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9** (1): 146.

Nado, J. 2014a. Philosophical Expertise. *Philosophy Compass*, **9** (9): 631-641.

———. 2014b. Why intuition? *Philosophy and Phenomenological Research*, **89** (1): 15-41.

———. 2015. Philosophical expertise and scientific expertise. *Philosophical Psychology*, **28** (7): 1026-1044.
Nersessian, N. 1992. In the theoretician's laboratory: thought experimenting as mental modeling. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **2** (291-301): 39.
Norton, J.D. 1991. Thought experiments in Einstein's work. In *Thought Experiments in Science and Philosophy*, T. Horowitz and Gerald J. Massey (eds.), Savage, MD: Rowman and Littlefield, 129-148.
———. 2004. Why Thought Experiments Do Not Transcend Empiricism. In *Contemporary Debates in the Philosophy of Science*, Christopher Hitchcock (ed.), Oxford: Blackwell, 44-66.
Oberle, C.D., M.K. McBeath, S.C. Madigan, and T.G. Sugar. 2005. The Galileo bias: A naive conceptual belief that influences people's perceptions and performance in a ball-dropping task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **31** (4): 643.
Pais, A. 1982. *Subtle is the Lord: The Science and the Life of Albert Einstein: The Science and the Life of Albert Einstein*. New York: Oxford University Press.
Sax, L.J., K.J. Lehman, R.S. Barthelemy, and G. Lim. 2016. Women in physics: A comparison to science, technology, engineering, and math education over four decades. *Physical Review Physics Education Research*, **12** (2): 020108.
Schulz, E., E.T. Cokely, and A. Feltz. 2011. Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition*, **20** (4): 1722-1731.
Schwitzgebel, E. and F. Cushman. 2012. Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, **27** (2): 135-153.
Sivasundaram, S. 2016. Interpreting Quantum Mechanics. *RePoSS: Research Publications on Science Studies*, **93**. http://css.au.dk/fileadmin/reposs/reposs-039.pdf.
Sklar, L. 1992. *Philosophy of physics*: Dimensions of Philosophy S.
Starmans, C. and O. Friedman. 2012. The folk conception of knowledge. *Cognition*, **124** (3): 272-283.
Stuart, M.T., Y. Fehige, and J.R. Brown. 2017. *The Routledge companion to thought experiments*. London: Routledge.
Swain, S., J. Alexander, and J.M. Weinberg. 2008. The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and phenomenological research*, **76** (1): 138-155.
Tobia, K., W. Buckwalter, and S. Stich. 2013. Moral intuitions: Are philosophers experts? *Philosophical Psychology*, **26** (5): 629-638.
Vaesen, K., M. Peterson, and B. Van Bezooijen. 2013. The reliability of armchair intuitions. *Metaphilosophy*, **44** (5): 559-578.

Weinberg, J.M., C. Gonnerman, C. Buckner, and J. Alexander. 2010. Are philosophers expert intuiters? *Philosophical Psychology*, **23** (3): 331-355.
Weinberg, J.M., S. Nichols, and S. Stich. 2001. Normativity and epistemic intuitions. *Philosophical Topics*, **29** (1/2): 429-460.
Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.
———. 2011. Philosophical expertise and the burden of proof. *Metaphilosophy*, **42** (3): 215-229.

Appendix 1

	PHYSICS	CONTROL	*p*
Background information	*N* = 57	*N* = 57	
Age (in years)	33 (10.02)	34 (6.72)	ns
Gender (female)	11 (19%)	27 (47.3%)	*
Holds a PhD	32 (56.1%)	33 (57.9%)	ns
Enrolled in a PhD programme	25 (43.9%)	24 (42.1%)	ns

Table 5: Background information about all participants and their education (p=.001).

Holds a PhD in:	*N* = 33
Anthropology	2
Chemistry	2
Economics	1
Engineering	1
History	4
Languages	1
Literature	1
Mathematics	1
Medicine	2
Musicology	1
Political science	11
Psychology	1
Philosophy	3
Sociology	2
Enrolled in a PhD programme	N = 24
Business & managements	2
Computer Science	2
History	1
Linguistics	2
Literature	1

Mathematics	4
Medicine	1
Political science	8
Psychology	2

Table 6: Information about the education of participants from the CONTROL group.

Specialization in PHYSICS group	*Number of participants who chose the answer*
Astrophysics and cosmology	9
Atomic- molecular- and optical physics	24
Biophysics	1
Solid-state- and materials physics	11
Sub-atomic physics	8
Nano physics	8
Statistical physics	5
Other*	11

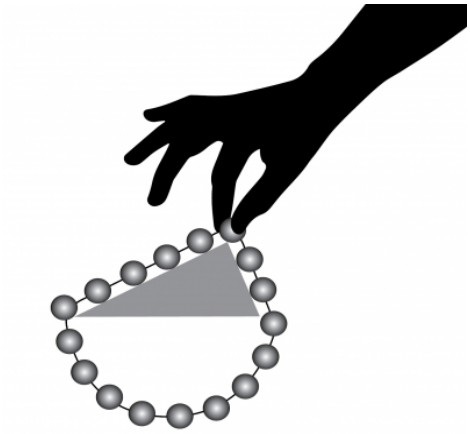
Table 7: Areas of specialisation of physicists via self-identification (multiple answers allows). Under "Other" subjects stated: Particle physics (3); Quantum physics (4); Fluid mechanics (1); Condensed matter physics (1); Nuclear physics (1) Applied physics and methodology (1).

Appendix 2

In our six tasks, we asked subjects to consider a scenario (S), answer a comprehension question (CS), and say whether they would agree with the judgements offered (J).

Stevin's chain

S: "Imagine that somebody put a chain with evenly spaced metal balls with the same size and weight on top of an inclined frictionless plane."



CS: "The inclined plane in the above scenario is …"

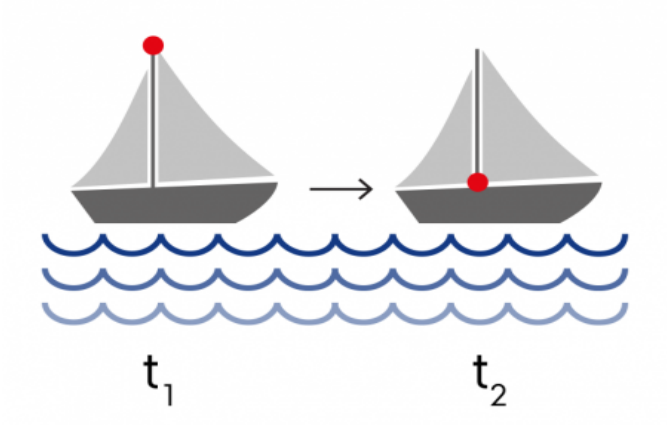
J: "Once the chain is released it will move sideways." [This is incorrect.]

J is the negation of the judgment elicited in a famous thought experiment by Simon Stevin. With this thought experiment Stevin wanted to demonstrate the plausibility of his claim that for inclined planes with the same height, the force needed to keep weights in their position on those planes varies inversely with the planes' lengths. More specifically, in the depicted scenario S Stevin used a pair of planes of which one was double the length of the other and the weights placed on the longer plane were double the amount of weights on the shorter plane. According to his law, the weights on those two planes (which are connected to each other) should balance each other out. In order to drive home the point, Stevin connected the weights on those two planes with a chain of further weights (seen at the bottom of the figure). Now, if one were to deny Stevin's 'law' and approve of the statement that the entire chain moves to the right or to the left (as in J), it's not clear how one could deny that the chain *keeps* moving either to the right or left. After all, the chain is uniform (equal weights, equal distances between the weights). But since this would

constitute a perpetual motion, which is ruled out by the 2nd law of thermodynamics. Ernst Mach, who discussed this task in his *The Science of Mechanics*, ruled this out on an "instinctive basis".

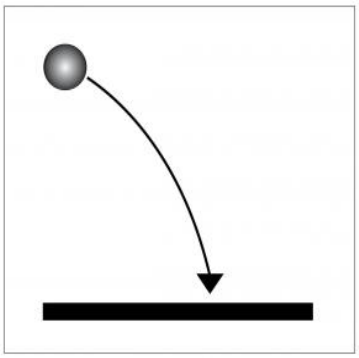
Galileo's ships

S: "Imagine yourself standing at the coast and observing a ship moving with constant speed. The picture shows a snapshot of the ship's movement at two points in time: t1 and t2. At t1, a cannon ball is dropped from the top of the mast of the ship and at t2 the cannon ball has reached its final position:"



CS: "As the observer you are located on …"

J: "When seen from the coast, the trajectory of the ball moving from t1 to t2 is as in the following picture:" [This is correct]



Galileo used this thought experiment in his *Dialogue Concerning the Two Chief Systems of the World* to persuade those believing in the geocentric system that a moving earth would not necessarily pose any problems for terrestrial physics (people were concerned that a moving earth would imply objects on earth flying through the air). The object falling from the top of the mast to its bottom on a moving ship illustrates that the trajectory of falling

objects may appear straight when it in fact decomposes into straight and rectilinear motion (as in our second picture). Galileo's ship also demonstrates what has come to be known as Galilean relativity: the classical laws of physics are the same in all inertial frames (and two inertial frames can be transformed into each other via Galilean transformations).

Newton's cannonball

S: "Imagine shooting a cannonball from a high elevation on earth into the distance. On the picture, you see the trajectories of a cannonball shot with (relatively) low speed, A, and with a higher speed, B. Cannon balls following A and B will land back on earth."

CS: "The cannonball following trajectory B will land on …"

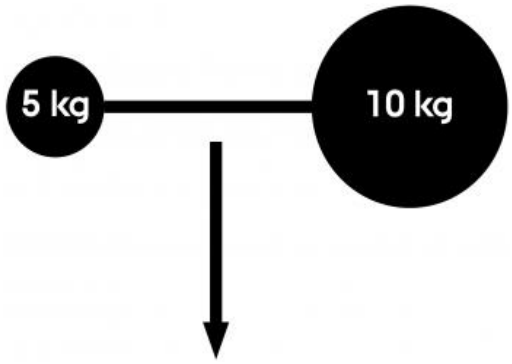
J: "Trajectory C is possible." [This is correct]

Newton used this thought experiment in the *The System of the World* to show that the orbital motion of the moon (and the planets around the sun) is accounted for by the same forces that act on earth (namely an inertial and a gravitational one).

Galileo's tower

S: "Imagine you connect a steel ball of 10kg and a steel ball of 5kg with a tight chain and drop the combined object from a high elevation in a vacuum. How does one determine the speed of fall of the combined object? One proposal is to average the speed of the two objects (when they fall separately): since 5kg falls slower than 10kg, the combined object

will fall slower than the 10kg ball. Another proposal is to add the weights: and since 15kg > 10kg, the combined object will fall quicker. Yet another proposal is that the combined object falls just as fast as the 10kg ball on its own, since the weight makes no difference to the speed of fall."



CS: "The combined object weighs …. kg."

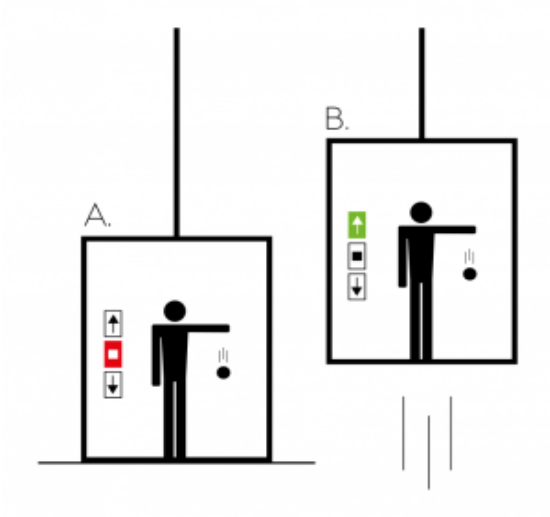
J: "The combined object will land just as fast on the ground as the 10 kg steel ball alone". [This is correct]

This is another thought experiment by Galileo, expounded in his *Dialogues concerning two new sciences*. Galileo used this thought experiment to demonstrate an internal contradiction in Aristotle's physics, according to which heavier bodies fall quicker to the ground than lighter ones: in situations such as the one described, Aristotelian physics implies a contradiction, namely that both the combined object falls quicker *and* slower than the heavy object alone. On the basis of this thought experiment (and other evidence), Galileo argued not only that Aristotelian physics is false, but also that all bodies fall at the same rate (which he could not demonstrate at the time, as he had no means for producing vacuums).

Einstein's Elevator

S: "Consider a person in the scenarios A and B. In A, the person is standing inside an elevator that sits on the ground level. In B, the person is inside an elevator that is dragged through empty space somewhere in the universe with uniform acceleration (i.e., the speed increases constantly). In neither A or B can the person see what's going on outside the

elevator. In B, the person does not feel that the elevator is being dragged: the elevator appears perfectly stable to her. Suppose that the person wants to find out whether she is in A or B by dropping a ball to the floor."



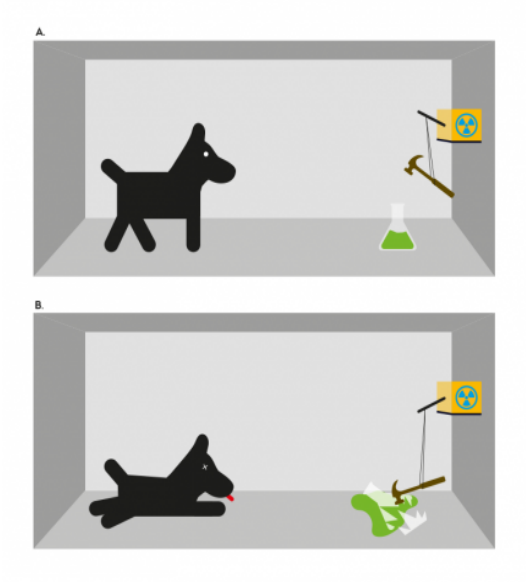
CS: "In B, the elevator is dragged through…"

J: "The person can determine whether she is in A or B by the manner in which the ball drops to the floor." [This is incorrect]

Einstein (and Infeld) used this thought experiment to illustrate the equivalence between inertial and gravitational forces, which underlies the general theory of relativity. The trajectories of the balls will only then be indistinguishable in the two scenarios if the acceleration equals the strength of gravity on the surface of the earth. This is suggested in the thought experiment by the person in the elevator "not feeling" any drag.

Schrödinger's cat

S: "Imagine a dog trapped in an opaque box. There is a very small amount of radioactive substance in the box: there is a probability of 50% that one atom of that substance decays within one hour. Whenever one atom of this substance decays, a Geiger counter will detect this atom and trigger the destruction of a flask containing a highly toxic substance. As soon as the flask breaks, the dog dies instantly. Suppose that the dog is kept in the box for one hour before the box is opened."



CS: "If one atom decays, then the dog will …"

J: "Before the box is opened the dog is either dead or alive." [Our expectation was that physicists should judge this as incorrect]

Erwin Schrödinger used this thought experiment to challenge Bohr and Heisenberg's Copenhagen interpretation of quantum mechanics.

The wave function of quantum mechanics describes the system in terms of probabilities. According to the Copenhagen interpretation, the probability of the state of a physical system at some point in time describes the actual system and is not just an expression of our own ignorance. The system is also said to be in a "superposition" of states. When we measure the system is said to "collapse" and the system adopts a definite state. Which state the system actually adopts upon measurement, however, cannot be determined within quantum mechanics.

Schroedinger's reasoning was that if the Copenhagen interpretation were correct, then the cat in the box (in our case: a dog) should be in a state of superposition before the opening of the box (the "measurement") causes a collapse of the wavefunction. However, since we would normally judge that the cat/dog does have a definite state before we open the box, the Copenhagen interpretation must be false. As explained in the main text, there

are legitimate ways of avoiding this conclusion. In our analysis we presumed that the correct response in this task would be the denial of J (but see above).

Appendix 3

Question-by-question chi-square tests in the STRICT condition for correctly answered questions by physicists vs. non-physicists.

	LAX		STRICT	
	X^2	*p*	X^2	*p*
Stevin's chain	X^2 (1, 114) =0.9522	.329	X^2 (1, 114) =3.5185	.061
Einstein's elevator	X^2 (1, 114) =0.3221	.570	X^2 (1, 114) =4.3804	.036*
Schroedinger's cat	X^2 (1, 114) =4.7281	.030*	X^2 (1, 114) =1.4902	.222
Newton's cannon	X^2 (1, 114) =3.0299	.082	X^2 (1, 114) =12.0689	.000***
Galileo's ship	X^2 (1, 114) =14.4190	.000***	X^2 (1, 114) =9.1200	.003**
Galileo's tower	X^2 (1, 114) =11.75257	.001**	X^2 (1, 114) =21.4023	.000***

Table 8: Chi-square tests in the STRICT and LAX conditions comparing physicists and non-physicists for each of our thought experiments. Statistically significant differences at the 95% level are marked with a star (*) at the 99% level with a double-star (**) and at the 99.9% level with a triple-star (***).

Appendix 4

Dependent Variable: Total Strictly Correct			
Variable	Model 1	Model 2	Model 3
Physics (dichotomous)	0.55*** (0.12)	0.49* (0.13)	0.41** (0.15)
Marginal Effect of Physics (How many more questions do physicists get right on average)	1.45***	1.34***	1.52** if female 1.22** if non-female
Physics * Female			0.27 (0.28)
Exposure to physics at university (but no bachelor degree) (dichotomous)		0.19 (0.25)	0.16 (0.25)
Age		0.004 (0.006)	0.005 (0.006)
Female		-0.31* (0.14)	-0.44* (0.20)
Natural log of duration (seconds)		-0.01 (0.07)	-0.004 (0.07)
Degree of English proficiency (1=intermediate, 2=advanced, 3=native)		-0.12 (0.12)	-0.12 (0.13)
Constant	0.670*** (0.09)	0.97*** (0.59)	0.99 (0.59)
Log Likelihood	-198	-194	-194
Ln_Alpha	-27.09	-27.09	-27.09
Alpha	1.71 e-12	1.71 e-12	1.71 e-12

Table 9: Negative binomial regression analysis for the STRICT condition. According to Model 1 the marginal effect of having a PhD degree in physics (or studying towards one) is 1.47. Thus, having a physics degree is predicted to increase the number of questions answered correctly by 1.47. Model 2 includes control variables for gender, exposure of controls to physics at university, age, duration of task performance, and level of English. Model 3 interacts the factor 'women' with physicists and non-physicists. It predicts that women with a PhD degree in physics (or ones studying towards one) judge 1.54 tasks more correctly than women in the control group. *denotes $p<0.05$ (95% statistically significant); ** denotes $p<0.01$ (99% statistically significant); *** denotes $p<0.001$ (99.9% statistically significant)

Dependent Variable: Total Lax Correct			
Variable	Model 1	Model 2	Model 3
Physics (dichotomous)	0.24** (0.10)	0.23* (0.11)	0.22 (0.13)
Marginal Effect of Physics (How many more questions do physicists get right on average)	0.93*	0.89*	0.79 if female 0.91 if non-female
Physics * Female			0.01 (0.23)
Exposure to physics at university (but no bachelor degree) (dichotomous)		0.23 (0.19)	0.23 (0.19)
Age		-0.002 (0.006)	-0.002 (0.006)
Female		-0.18 (0.11)	-0.19 (0.15)
Natural log of duration (seconds)		0.04 (0.06)	0.04 (0.06)
Degree of English proficiency (1=intermediate, 2=advanced, 3=native)		-0.03 (0.10)	-0.03 (0.10)
Constant	0.21*** (0.07)	1.14* (0.49)	1.14* (0.49)
Log Likelihood	-203	-201	-201
Ln_Alpha	-21.47	-21.47	-21.47
Alpha	4.75 e-10	4.75 e-10	4.75 e-10

Table 10: Negative binomial regression analysis for the LAX condition. According to Model 1 the marginal effect of having a PhD degree in physics (or studying towards one) is 0.96. Model 2 includes control variables for gender, exposure of controls to physics at university, age, duration of task performance, and level of English. Model 3 interacts the factor 'women' with physicists and non-physicists. It predicts that women with a PhD degree in physics (or ones studying towards one) judge .83 tasks more correctly than women in the control group. *denotes $p<0.05$ (95% statistically significant); ** denotes $p<0.01$ (99% statistically significant); *** denotes $p<0.001$ (99.9% statistically significant)