Synthese (2022) 200:221
https://doi.org/10.1007/s11229-022-03637-7

ORIGINAL RESEARCH

Can thought experiments solve problems of personal identity?

Lukas J. Meier[1]

Received: 30 August 2021 / Accepted: 23 February 2022 / Published online: 12 May 2022

> The method of science fiction has its uses in philosophy, but (...) I wonder whether the limits of the method are properly heeded. To seek what is 'logically required' for sameness of person under unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with.
>
> (W. V. O. Quine)

Abstract
Good physical experiments conform to the basic methodological standards of experimental design: they are objective, reliable, and valid. But is this also true of thought experiments? Especially problems of personal identity have engendered hypothetical scenarios that are very distant from the actual world. These imagined situations have been conspicuously ineffective at resolving conflicting intuitions and deciding between the different accounts of personal identity. Using prominent examples from the literature, I argue that this is due to many of these thought experiments not adhering to the methodological standards that guide experimental design in nearly all other disciplines. I also show how empirically unwarranted background assumptions about human physiology render some of the hypothetical scenarios that are employed in the debate about personal identity highly misleading.

Keywords Objectivity · Personal identity · Reliability · Scientific method · Thought experiments · Validity

1 Introduction

Thought experiments are mental test scenarios that purport to deliver scientifically acceptable results in the absence of actual physical execution. Scientists use imaginary situations as a method to test hypotheses, to expose contradictions, or to explore the scope of concepts (Brownlee & Stemplowska, 2017, p. 21). Hypothetical reasoning is

✉ Lukas J. Meier
ljm204@cam.ac.uk

[1] Churchill College, University of Cambridge, Cambridge CB3 0DS, UK

employed in a variety of disciplines, including in physics and economics, and it has a particularly long and important tradition in philosophical discourse, which began as early as in pre-Socratic times.[1] Also engaging in a meta-discourse on the thought-experimental technique, however, is a comparatively recent phenomenon.

The term *thought experiment* was introduced in 1811 by Danish physicist and philosopher Hans Christian Ørsted (1998, p. 296) and later became popular through the writings of Ernst Mach (1926), who was the first systematically to consider thought experimentation as a scientific method. The past thirty years finally saw heightened interest in the topic, culminating in the publication of several monographs (Bertram, 2012; Cohnitz, 2006; Gendler, 2000; Häggqvist, 1996; Kühne, 2005; Rescher, 2005; Sorensen, 1992).

The thought-experimental method has had many prominent advocates, including such major figures like Descartes and Leibniz. Proponents of thought experimentation usually maintain that

> philosophy is the analysis or articulation of the conditions of application of our concepts. As masters of these concepts (…) we have at least an implicit grasp of their application conditions; this tacit knowledge of when they apply and when they should be withheld can be manifested equally well in real and imaginary cases (Johnston, 2016, p. 91).[2]

Appealing to intuitions about imaginary cases has also seemed dubious to some, however. At the beginning of the twentieth century, French physicist Pierre Duhem (1906, p. 331) made the following observation.

> Invoquer une telle expérience fictive, c'est donner une expérience à faire pour une expérience faite; c'est justifier un principe non pas au moyen de faits observés, mais de faits dont on prédit la réalisation; et cette prédiction n'a d'autre fondement que la croyance au principe à l'appui duquel on l'invoque.[3]

Some contemporary authors echo this position. Ulrich Kühne (2005, p. 10) asserted that thought experiments are experiments of which the main part is missing;[4] and Bernard Williams (1970, p. 179 f.) worried that it is often the way in which an author *describes* a certain hypothetical situation that determines whether or not it appears to support a particular theory, while a slightly different account of the same setting could yield entirely different results. Is this criticism well-founded?

Especially in debates about personal identity, philosophers have been relying heavily on a 'seemingly endless litany of fantastical thought experiments', and the intuitions

[1] See Rescher (2005, pp. 61–72) for some examples from that period. Probably the most famous ancient thought experiment is Plato's Tale of the Ring of Gyges (*Republic*, 359d–360a). Plato (1997, p. 1000) inquired whether one would remain moral if all sanctions were removed, which he tried to establish by imagining that there existed two rings that made their owners invisible, one worn by a just, the other by an unjust person.

[2] See also Noonan (2003, p. 199 f.).

[3] Translation (by L.J.M.): 'Employing such a hypothetical experiment is passing off an experiment yet to be executed as one already performed; it is justifying a principle not on the basis of observed facts, but on the basis of facts whose realisation one predicts; and this prediction has no other foundation than the belief in the principle on the basis of which one postulates it.'

[4] See also DeGrazia (2005, p. 26 f.) and Norton (1996, p. 335).

that these hypothetical situations elicit serve as weighty evidence in favour or against the proposed accounts and concepts (Blatti, 2019). One of the main reasons for the great dependence on thought experimentation in this field is that deciding between the two main approaches to the question of what we essentially are—biological and psychological accounts—requires situations in which bodily and mental characteristics come apart. In real-life settings, an individual's bodily continuity and the continuity of his or her mental features either occur conjoined or else bodily continuity occurs in isolation, as in a persistent vegetative state. While we cannot learn much from the former case, we do not know how to interpret the latter. The interesting permutation, it appears, is the third one: the presence of psychological features in the absence of bodily continuity. Especially proponents of psychological accounts of personal identity therefore often introduce hypothetical situations that are designed to provide us with this configuration, for the study of which we cannot resort to empirical evidence.

Pioneered by John Locke's case of the prince whose soul enters a cobbler's body and his thought experiment featuring the rational parrot (Locke, 2008, II.XXVII, § 15 and § 8), authors have made frequent use of a great variety of hypothetical situations to prove or disprove their respective views about personal identity. We are invited to envisage being teletransported to Mars (Parfit, 1987, p. 199), existing as mere brains in vats (Putnam, 2000), or even being on a mission to retrieve a 'Supersonic Tunneling Underground Device' whose special type of radiation makes it necessary to remove the brain and connect it to the decerebrated body by means of 'microminiaturized radio transceivers' (Dennett, 1998, p. 310 f.). Such thought experiments are certainly very creative. But are they also suited to act as testing grounds for hypotheses concerning our synchronic and diachronic persistence? What epistemic status can one grant the results that this method delivers?

In this paper, I shall describe what I take to be the two most severe methodological weaknesses of using thought experimentation to solve problems of personal identity, and I shall strictly limit the focus to this very area of philosophy. I will be arguing that since questions of personal identity often require hypothetical scenarios that are very distant from the actual world, many of the latter do not comply with the standards of good design that commonly guide physical experimentation, which is why they are ineffective at resolving conflicting intuitions. I shall also submit that these scenarios can be misleading as their authors tend to make empirically unwarranted background assumptions about human physiology. I will illustrate each claim with well-known examples from the literature.

2 Nonconformity to the standards of scientific experimental design

Advocates of biological accounts and proponents of psychological views of personal identity differ in what they claim their respective intuitions are about situations in which bodily and mental features come apart. The former, believing that we are identical with our living organisms, find it perfectly natural to suppose that we are wherever our bodies are located. Animalism, the most prominent variant of this type of views, has recently been attracting a growing number of proponents (Blatti, 2012; Blatti &

Snowdon, 2016; DeGrazia, 2005; Inwagen, 1995; Merricks, 2001; Olson, 1997, 1999; Snowdon, 1990; Wiggins, 1980, 2003).

Conversely, advocates of psychological accounts hold that our persistence must consist in the continuity of some kind of mental relation. According to this view, each of us was the past being whose mental features he or she has inherited; and he or she will be the future being who is equipped with these mental features. There is much disagreement, however, over what exactly these psychological characteristics may be. The most influential view is due to John Locke (2008, II.XXVII, § 9), who suggested that memories form the decisive relation. Modified versions of this traditional account continue to have many prominent supporters (Green & Wikler, 1980; Lewis, 1976; Parfit, 1971, 1987; Perry, 1972; Shoemaker, 1970, 1999, 2008; Shoemaker & Swinburne, 1984).

Which of the two mutually exclusive positions is preferable is often to be established by constructing hypothetical test scenarios in which physical and mental characteristics come apart. Usually, the opponents reach differing conclusions even when considering one and the same thought experiment (Nichols & Bruno, 2010, p. 297; Sider, 2001, p. 197); improved imagined situations are then devised (Talbot, 2013, p. 332), but instead of settling the matter, they often only cement the disagreement (Cohnitz, 2006, p. 165; Johnston, 2016, p. 96; Norton, 1996, p. 361). Seldom does any newly suggested thought experiment manage to put the respective issue to rest.

How can the two camps disagree about the conclusion derived from the very same thought experiment? In physical experiments, as they are conducted in the natural sciences, the hypothesis that is being tested can normally be regarded as either confirmed or refuted when the experiment was carefully designed and carried out according to appropriate standards. Not so in philosophy. More than three centuries have passed since Locke first introduced his thought experiments into the debate. If one does not doubt that there *is* something to discover in questions of personal identity, one may start to question the aptness of employing thought experiments as the dominant scientific method in this area of philosophy. What is it about hypothetical situations that enables them to evoke so radically differing reactions?

The answer may lie in the fact that solving problems of personal identity often requires thought experiments so fantastical that they transgress the standards to which physical experiments are commonly held. Scientists demand of physical experiments that they be objective, reliable, and valid. Put simply, an experiment is *objective* if it manages to exclude all unwanted outside influences on its result; it is *reliable* if, whenever repeated, it always yields the same outcome; and it is *valid* if it measures what it claims to be measuring. These are widely agreed standards in the scientific community (Lienert & Raatz, 1994, pp. 7–14; Nelson, 1980; Schwartz-Shea & Yanow, 2012, pp. 92–98; Tetens, 2016).

Why should the standards of physical experimentation be relevant to thought experimentation at all? Is experimenting with thoughts not a different domain, so that what applies in one case does not necessarily apply in the other? For many thought experiments in philosophy, the factual 'scaffolding' on which the imagined situation relies has been independently empirically established: we already know, for example, how railroad switches work to divert trolleys in cases that test our ethical judgments (Thomson, 1976); and we know what it means to have ten coins in one's pocket—as in

Edmund Gettier's (1963, p. 122) famous example in modern epistemology. These thought experiments consist nearly entirely of *ceteris-paribus* conditions, which do not exceed what can be regarded as well-established knowledge about the world. All required empirical work has been carried out already, long before philosophical considerations enter the picture. Hence, when we apply our philosophical theories and concepts to these cases, we really *only* test our intuitions. The standards that guide experimental design in the sciences are therefore not decisive.

Thought experiments in the domain of personal identity are often quite different. Many of them rely, as we shall see, on assumptions about the consequences of drastic changes to what is the case in the actual world—like on the hypothesis that whole brains are divisible and can be transplanted into different bodies (Parfit, 1987, p. 254 f.) or on the idea of people splitting like amoeba (Williams, 1973, p. 23). If the imagined scenario departs from the empirical data that we possess about the external world, the thought experiment becomes akin to an empirical experiment in its own right. Its scope now greatly exceeds the mostly conceptual question that the thought experiment was intended to raise. As I shall show, these modified empirical background assumptions are not simply inert illustrative embellishments, but factors directly relevant to the intuitions that the imagined setup elicits and thus to the conclusion that the thought experimenter may derive from it.

'If we have to do with a real thought experiment, the empirical data upon which it rests must have been both well-known and generally accepted before the experiment was even conceived', warned Thomas Kuhn (1978, p. 241). To acquire such an empirical basis for one's intuitions, it would therefore be best first to conduct a physical experiment. However, thought experiments in personal identity often do not permit the verification of their empirical premises: in the case of brain bisection for ethical reasons, in that of humans dividing like amoeba for metaphysical, and in Derek Parfit's famous teletransportation case for technological ones.

Philosophers will be quick to point out that actual physical execution may not always be required—not even when empirical facts are at stake. They may, for instance, refer to Albert Einstein's (1905, p. 891) thought experiments that ultimately led to the development of special relativity or to Erwin Schrödinger's (1935, p. 812) cat paradox. Whether any novel *empirical* knowledge can be derived from hypothetical setups is a fascinating question that we must, however, leave aside.[5] What seems less controversial is that *if* hypothetical reasoning could indeed sometimes replace physical execution, it would likewise have to conform to the standards that guide such experimentation. We know that scientific experiments conducted in the external world are prone to having inaccurate outcomes when they are not objective, reliable, and valid. How, then, could the same scientific questions, when posed to the mind instead of to the external world, yield correct results if the corresponding thought experiments are not likewise executed in accordance with these principles? As Ernest Sosa (2007, p. 106) put it: 'The way intuition is supposed to function in epistemology and in philosophy more generally (…) is by analogy with the way observation is supposed to function in empirical science'.

[5] See the debate between Brown (2004) and Norton (1996), as well as Kuhn (1978).

The first part of the paper will therefore be concerned with showing why the bizarre nature of many thought experiments in the domain of personal identity prevents scientific accuracy: while their fantastical setup means that some of their yet unverified empirical premises would need evaluation—which necessitates observing the aforementioned standards—it is also this very fancifulness that *precludes* the adherence to these paradigms. And devoid of factual certainty, I shall argue, whatever intuitions we may form about the imaginary situations are unfounded.

One may retort, with Parfit (1995, p. 15), that some thought experiments in the debate about personal identity are meant to discover 'not what the truth is, but what we believe'.[6] Thus, one might assert that the standards of good experimental design would not need to be followed since empirical facts are not what is at issue.[7] However, authors who are convinced that there is a fact to the matter as to what we essentially are and in what our synchronic and diachronic persistence consists must disagree with Parfit's premise: whether we would, for example, survive procedures like brain transplantation, they would hold, is not something that primarily depends on what we believe but on facts about the world—in this case on the neurological and physiological characteristics of our bodies and brains. This would locate the subject at least partly in the empirical realm.

Moreover, even if one maintains that questions of personal identity are purely conceptual issues to which our intuitions are the best guides, it is still true that what we believe depends in large parts on *what is the case*. The great disunity in response to the more fantastical thought experiments precisely shows that often we do not know what to believe. This is so, I shall argue, because these imagined situations rest on empirically shaky grounds—grounds that would have to, but cannot, be solidified by conducting (thought) experiments adhering to the standards of good experimental design. I will now illustrate each of the three principles with an example and try to establish whether classic thought experiments from the literature conform to them.

2.1 Objectivity

When one wants to establish one's weight, one steps onto the bathroom scales. The outcome of this very simple experiment is objective if the value that the scales display is the result only of one's body mass and the magnitude of the local gravitational acceleration. Other factors, like the room temperature or one's political views, must not be taken into account.

Most experiments are much more complex than this simple model, and the possible sources of error that jeopardise their objectivity are abundant. An important prerequisite of ensuring that the result of an experiment is objective is therefore strictly to differentiate between *causality* and *correlation*, that is, to distinguish a change in outcome that results from a modification of the factor under consideration from one that simply occurs contemporaneously. Hence, in experiments in the natural sciences only very few variables are actively being manipulated while all the remaining ones are

[6] See also Parfit (1987, p. 200).

[7] Thanks to an anonymous reviewer for raising this objection.

held as constant as possible. This is what distinguishes a proper experiment from mere observation, where the surrounding conditions are not under the scientist's control.

The parameter that is being influenced is called the *independent variable*; the one that is monitored for change is the *dependent variable*. If only one variable is being manipulated at a time, any difference in outcome is attributable to this very change alone. When an uncontrolled factor emerges, however, the experiment's objectivity is threatened as it is now unclear whether the change in the dependent variable is indeed caused by the manipulated independent variable (causality) or whether it results from a different source (mere correlation). Consequently, if a team of scientists want to determine which of the two drugs that they have developed relieves pain, they either form two groups of patients, of which one receives drug *x* while the other group is treated with drug *y*, or they give both drugs to the same group at different times. What they must not do, however, is to administer both substances to the same patients at the same time as this simultaneous modification of two crucial variables would render the result unattributable to the variable that is actually causally responsible. The experiment's designers also need to control for the placebo effect, for the test persons' varying bodily dispositions, and for other factors that are known to influence pharmaceutical studies. Once they have correctly factored in all potential influences, the experiment should be objective. Is the same true of hypothetical experiments in personal identity?

In his defense of the thought-experimental method, Daniel Kolak (1993, p. 46) maintains that 'in thought experiments about persons, a properly imagined set-up allows us to leave out all factors but the one under examination'. If this were indeed the case, the objectivity of thought experiments would exceed that of most physical experiments, in which, despite meticulous planning, it is often impossible to control each and every variable. If thought experimentation was superior in this regard, one should expect great unanimity among philosophers: unless there was a problem with the experiment's reliability or validity, everyone should happily accept the outcome. This is not as things stand. Let us examine why this is so on the basis of what is probably the most famous thought experiment in the literature about personal identity: Derek Parfit's teletransportation case.

> I enter the Teletransporter. (…) When I press the button, I shall lose consciousness, and then wake up at what seems a moment later. In fact I shall have been unconscious for about an hour. The Scanner here on Earth will destroy my brain and body, while recording the exact states of all of my cells. It will then transmit this information by radio. Travelling at the speed of light, the message will take three minutes to reach the Replicator on Mars. This will then create, out of new matter, a brain and body exactly like mine. It will be in this body that I shall wake up (Parfit, 1987, p. 199).[8]

The imagined situation is intriguing and it has sparked off a long and lively debate. Will it really be *I* who wakes up in the newly created body? Parfit confronts us with a world that is radically different from our actual one—even more so than it would initially seem. It is a world in which technology is so advanced that there exists a machine

[8] For a slightly different setup, see Carruthers (1995, p. 198).

that can translate the characteristics of living matter into information to manufacture a qualitatively identical duplicate according to the blueprint. Such a procedure may not be possible in the actual world, but that is not the main worry. In thought experiments we are precisely invited to speculate, and the stage of technological advancement, one may argue, is irrelevant to the question of personal identity. The problem is that with a world so disposed other factors creep in—further independent variables—that are not explicitly mentioned in the description but have the potential to influence the conclusion (Wilkes, 1999, p. 45).

One should, for example, expect the inhabitants of a world in which teletransportation devices are available to have very different attitudes towards life and death. A person's life would be something that could be suspended and stored on any data medium to become later manifested in different matter. As there is no requirement that the blueprint, once created, is immediately transmitted to the replicator on Mars, this imagined world must also include the option of time travel into the future. Before entering the teletransporter, one could instruct the machine to delay the transmission for hundred years and finally leave the replication booth without having grown older even a single day.[9] It would be a world in which parents could meet their children at an older age than they have reached themselves if, after a child is born, the parents enter the teletransporter and delay replication for a sufficient amount of time. In what would a person's death in such an environment even consist? Would it be the deletion of the final remaining copy of the blueprint? Or would it be the ceasing to exist of the last living clone, on the condition that another replication will never be attempted?

Phrased in the terminology of physical experiments, these are independent variables, that is, manipulable integral parts of the setting in which the experiment takes place that potentially exert an influence on its outcome. Instead of, as it initially appears, making alterations only to the one specifically named parameter that distinguishes the imagined world from our actual one—'humans can be teletransported'—a great number of other variables are also tacitly amended. Among these are so elementary ones as 'adult human beings can be created from non-living matter' and 'people can travel to the future'.

With a large number of independent variables modified at the same time, it becomes impossible to determine whether the observed change in the dependent variable is indeed the result only of the one condition of the actual world that the thought experiment was originally supposed to waive or just *correlates* with it, while stemming from any of the other independent variables that were also modified. Causality and correlation cannot be distinguished.

Unlike Locke's original scenario in which the prince and the cobbler exchange souls—which raises an even greater number of methodological questions—Parfit employed his teletransportation thought experiment for much wider purposes than merely teasing apart physical and psychological continuities. Famously, he drew from it ingenious conclusions about causation as well as about the nature and significance of identity relations (Parfit, 1987, Part 3). Thus, I do not mean to suggest that one cannot learn anything important from this hypothetical situation—quite the reverse.

[9] Whether this applies to numerically or qualitatively identical individuals depends, of course, on the conclusion that one derives from the thought experiment.

The question is, rather, to what extent the conclusions arrived at are applicable to *our* persistence, given that the background before which the thought experiment takes place is so much unlike our actual world.

To be clear, physical experiments, too, often struggle to conform to the ideal of objectivity. Eliminating all unwanted influences is a very complicated undertaking, not only in the domain of hypothetical experimentation. Rather than identifying a categorical difference, one may therefore locate physical experiments and thought experimentation on a continuum with regard to the parameter of objectivity.

There is, however, one important dissimilarity: when natural scientists are unable to banish all undesired influences in physical experiments, certain countermeasures are available. These include the use of positive and negative controls, blinding, randomisation, and other statistical devices. If implemented correctly, such measures are very powerful. Can unwanted influences in thought experiments be balanced out in the same ways?

For the past two decades, a movement termed *Experimental Philosophy* has been setting out to do exactly this. Most of its proponents share with traditional armchair analytic philosophy the basic assumption that intuitions provide a trustworthy source of evidence for philosophical investigation; however, they differ from the former in what they regard as the proper way of obtaining this data. While traditional armchair philosophising deems each individual's own intuitions sufficiently representative, the evidence on which experimental philosophers rely is mostly survey data (Alexander, 2010, p. 297 f., 305; Sosa, 2007, p. 100; Weinberg & Crowley, 2009, p. 227). Thus, this branch of philosophy focuses on 'many of the same types of intuitions that have long been at the center of philosophical study, but it examines those intuitions using the methods associated with contemporary cognitive science—systematic experimentation and statistical analysis' (Knobe, 2007, p. 81). Survey methods have been applied across a range of fields, including the philosophy of action, language, and mind, as well as ethics and epistemology (Buckwalter & Turri, 2018, pp. 282–291; Ludwig, 2007, p. 128). A few studies even investigated people's intuitions regarding questions of personal identity (Nichols & Bruno, 2010; Strohminger & Nichols, 2015; Tobia, 2015). This way, experimental philosophers aim to uncover—for instance—cultural, social, and educational variations in intuitions about philosophically significant questions (Buckwalter & Turri, 2018, pp. 285–287; DePaul & Ramsey, 1998; Hannon, 2018, p. 4148; Higgins & Dyschkant, 2014, p. 381; Ichikawa, 2013, p. 47; Ludwig, 2007, p. 128; Nisbett et al., 2001; Weinberg et al., 2001).

Experimental Philosophy rests on the conviction that, if empirically obtained, intuitions constitute a robust source of philosophical evidence (Alexander, 2010, p. 298; Hannon, 2018, p. 4159). As will become clear in the course of this paper, I agree with the advocates of this movement that philosophy should make use of much more empirical data than it currently does (Ichikawa, 2013, p. 51); I disagree, however, that intuitions elicited by fantastical thought experiments—irrespective of whether acquired first-personally via introspection or third-personally with the help of surveys—are objective and reliable.

Multiple studies have shown that 'intuitions vary according to factors irrelevant to the issues thought-experiments are designed to address', so that intuition seems to be 'an unacceptably shifting foundation' (Swain et al., 2008, p. 153 f.). Intuitions

are sensitive to the context in which one considers thought experiments (Swain et al., 2008; Talbot, 2013, p. 330; Weinberg & Crowley, 2009, p. 229), to the direction in which change occurs in the imagined scenarios (Tobia, 2015) as well as to the presence or the absence of specific elements (Nichols & Knobe, 2007; Uhlmann et al., 2009).[10] As I shall argue in the subsequent section, they also vary according to *who* conducts the thought experiment.

Unlike in physical experimentation, it therefore does not seem to be helpful to form control groups of thought experimenters, to blind the experimenter to some facts of the hypothetical scenario, or to apply statistical methods to the imagined situation. Consequently, in thought experiments that are as distant from reality as most setups in the debate about personal identity, the tacit introduction of spurious variables is not only very common; the countermeasures used in physical experimentation also appear to be impotent in neutralising them.

2.2 Reliability

The second principle of proper experimental design is reliability. Going back to our initial example of establishing one's weight, the experiment is reliable when the scales always display the same result unless there is an actual change in body mass. When one steps on and off the scales a hundred times in a row and the established value remains constant, the experiment is most likely reliable.[11] The constancy of measurements under unchanged conditions is crucial to ensure that experiments are comparable and that their results are reproducible.

How reliable is experimenting with hypothetical situations? Intrapersonally, that is, when one and the same individual repeats a thought experiment, the results that it delivers are often rather consistent. While there have been cases in which philosophers have changed their minds about certain imaginary situations,[12] they usually adhere to their favourite interpretation.

However, scientific experiments count as producing reliable results only if the latter are *inter*personally consistent, that is, if *different* experimenters come to the same conclusion. Whether, for instance, a drug to be tested is administered by a doctor who appears trustworthy or by a colleague who gives off an unskilled impression may influence how patients rate the effectiveness of the medication that they receive. Well-devised experimental designs consequently aim to minimise these effects to the largest possible degree to achieve intersubjective reliability. Regardless of whether the same person conducts the experiment twice or different individuals execute it independently of each other, the result should not change.

[10] But compare Liao (2008).

[11] In this context it is indecisive whether the experiment indeed measures what it purports to be measuring—this is a question of *validity* (see below). If, for example, the result displayed is consistently 1 kg lower than the individual's true weight, the test is still reliable; it is just not valid.

[12] One might argue that in these cases the thought experimenters have arrived at their new positions in response to persuasive philosophical debate rather than through re-evaluating the imaginary scenario. If true, however, this only shows that hypothetical experiments are weak decision factors that can be overruled by other considerations if deemed appropriate.

Not only thought experiments in the domain of personal identity have evoked highly contradictory interpretations from different individuals; many famous hypothetical situations employed in other areas of philosophy, like Searle's (1980) *Chinese Room* or Jackson's (1982) *Mary the super-scientist*, share this fate.[13] Several studies in Experimental Philosophy uncovered substantial discrepancies when considering one and the same imaginary scenario (Machery et al., 2004; Nisbett et al., 2001; Weinberg et al., 2001), revealing 'significant (and surprising) inter- and intra-personal intuitional instability' (Alexander, 2010, p. 298).

Thought experimenters can, of course, share their intuitions through the medium of language; we talk about what is on our minds and, via this indirect route, exchange the intuitions that hypothecial situations elicit in each of us. Take, for instance, Galileo's thought experiment with which he sought to disprove the Aristotelian theory of gravity (Galilei, 1974, pp. 65–67). The fact that each individual executes the experiment only in his or her head does not present a difficulty as the setup is so straightforward that one can expect great similarity between the scenarios pictured in people's minds. After careful discussion, the thought experimenters would likely reach the same verdict—just as observers of the experiment's physical execution would presumably have come to a unanimous conclusion if Galileo had indeed dropped two connected objects from the Leaning Tower of Pisa.

However, the gap between individual thought experimenters widens when the imagined scenarios are more distant from the actual world. As detailed in the foregoing section, usually only a few parameters of hypothetical situations are made explicit in the description that their respective authors give, while many are left unarticulated. The subjects who carry out the experiments must therefore deal, consciously or subconsciously, with a multitude of indeterminate variables. The more the hypothetical situation differs from reality, the greater becomes the number of variables that demand specification, yet the fewer background conditions are automatically fixed by reference to the world as we know it—unlike in physical experimentation. As long as the thought experimenters recognise the modifications, language still enables their verbal comparison. But in imagined scenarios as distant as, for example, teletransportation, the number of alterations that the imagined world requires vis-à-vis the actual one becomes so large that thought experimenters can grasp the radically different possible world only intuitively rather than possessing an exhaustive list of counterfactual propositions. Many implicitly made assumptions will therefore escape introspection (Talbot, 2013, pp. 318, 321, 325), which, in consequence, prevents their verbalisation and intersubjective comparison. The human mind is, from a third-personal perspective, a less accessible testing ground for scientific hypotheses than the external world, which is open to all observers alike. Consequently, one should expect fanciful thought experiments to yield wildly contrasting verdicts, which is indeed what we observe. While this is a typical feature of the thought experiments employed in the debate about personal identity, the problem of restricted intersubjective comparability pertains to *all* thought experimentation that relies on fantastical situations.

[13] An overview of the difficulties with these thought experiments can be found in Cole (2020, Sect. 4) and in Nida-Rümelin & O Conaill (2019, Sect. 4).

As in the case of objectivity, certain procedures are in place to identify and correct poor reliability in physical experiments. The easiest strategy is to repeat the test, if necessary several times, with different sets of experimenters while other factors are being held constant (Tetens, 2016, pp. 42–45). When experimenting in thought with distant possible worlds, however, exchanging the experimenter *and* conducting the experiment under otherwise identical conditions is not feasible. How one spells out the many background assumptions that are fixed neither by the thought experiment's initial description nor by reference to the actual world is highly subject-specific: it depends on the thought experimenter's general philosophical beliefs, on his or her cultural background, and on many other individual characteristics. Since some of these are not available to introspective evaluation in an explicit form, exchanging the experimenter would inevitably mean also to replace most of these background assumptions. Isolating the subject conducting the experiment from the object being studied, as would be required for proper intersubjective control and thus for achieving reliability, is therefore impossible.

2.3 Validity

The three principles of good experimental design form a certain hierarchy. Objectivity is a necessary, but not a sufficient, condition of reliability; and reliability is a necessary, but not a sufficient, condition of validity. Hence, an experiment that is not objective can be neither reliable nor valid. An experiment that is not reliable may, in turn, well be objective but it cannot be valid either. I have shown that many thought experiments employed in the debate about personal identity do not fulfil the requirements of objectivity and reliability. If this is correct, their validity is equally threatened.

An experiment is valid when it measures the very parameter that it was designed to measure. If the scales in our example determined, for instance, the room temperature, this could well be an objective measurement (if it was free from other influencing factors) and it could also be a reliable one (if the established value always corresponded to the true temperature), but the experiment would not deliver a verdict on the parameter demanded of it, namely, the person's weight. Consequently, it would not be valid. Now consider the following hypothetical situation due to Peter Unger (1990, p. 205).

> One of my half-brains may be gradually both bisected and fitted with radio transceivers at the opening interface. To get a very gradual spectrum, we may use this plodding procedure: At each stage, we always bisect just one largest brain-part of those then available in the situation. So, after we have half-brains, we will have one half-brain communicating with two quarter-brains. Then there will be four quarter-brains communicating; then two eighth-brains and three quarter-brains all communicating; then four eighth-brains and two quarter-brains, and so on, and so forth. This bisecting procedure can be repeated time and again, with arbitrary assignment to one side or the other in cases where the starting number of cells is odd, not cooperatively even. Eventually each neuron of my present brain will be in a supportive dish of its own, in splendid isolation from the others, while

> hooked up to an enormously complex device that is, among other things, a radio communicator. At any stage in this spectrum of radio communications, each resulting brain-part may be moved so that it is a few miles from the others then maximally intact. (…) Will I exist in such a tremendously scattered condition as that?

To be valid, the thought experiment should demonstrate how our stream of consciousness would behave in a possible world that is mostly like the actual world, but contains a minimal technological alteration that is needed to make the experiment work. What it in fact establishes, however, is whether we would presumably continue to exist when a machine replaced the brain's synaptic circuitry while the neurons themselves remained organic, when cerebral tissues could be sliced into infinitely small units that nonetheless retained all of their functions, when data could be extracted from such minute portions of brain tissue and be transmitted to other units in real time, and so forth. What can we learn about ourselves from hypothetical situations of this kind?

With so many conditions of the actual world being manipulated, it becomes rather unclear whether the experiment really measures what it is supposed to measure, and its validity is therefore severely endangered. After all, if one put the scales from our example on the moon, the displayed weight would be different, too, although the change would just be one of location rather than a modification of the inner workings of our brains. Thus, while we can imagine scenarios like Unger's and formulate questions about these possible worlds, it is very doubtful that the answers we might find in them should tell us what we had originally sought to establish about *our* world and *our* persistence. Thompson and Cosmelli (2011, p. 174) remark that

> if all that matters is conceivability, then we can avail ourselves of whatever conceivable technical resources we need, regardless of whether such resources are remotely feasible or even possible in our world or in worlds with our laws of nature. But such conceivability or possibility in principle tells us virtually nothing of interest with regard to what concerns us here, namely, the explanatory framework of the neuroscience of consciousness.[14]

However, many hypothetical situations on which philosophers rely to probe their intuitions are exactly that: testing grounds so distant that the results they yield can hardly be pertinent to our world as it is, and thus be of any relevance to the hypothesis that the thought experiment is meant to evaluate or to the concept that it seeks to explore (Wittgenstein, 1967, p. 64). Hence, they are not valid.

3 Unwarranted assumptions about human physiology

I have been arguing that many thought experiments in the debate about personal identity fail to conform to the basic methodological standards of scientific experimentation, which makes them ineffective at resolving conflicting intuitions. This, I submitted, is mainly so because the imagined *possible* worlds are often bizarre. A further worry is this: some thought experiments employed in questions of personal identity actively

[14] See also Wilkes (1999, p. 46). Beck (2014, p. 193) does not find this problematic.

mislead by making unwarranted background assumptions about physiological facts obtaining in the *actual* world.

Let us consider another prominent example. Probably the most frequently used hypothetical scenario in the debate is that of brain transplantation. Sydney Shoemaker was the first to present such a case, and several authors have been following his lead, suggesting many different variations (Perry, 1972, p. 463; Williams, 1970, p. 162 f.). They are modern versions of the Lockean scenario in which the prince and the cobbler switch souls.

> First, suppose that medical science has developed a technique whereby a surgeon can completely remove a person's brain from his head, examine or operate on it, and then put it back in his skull (regrafting the nerves, blood vessels, and so forth) without causing death or permanent injury; (…) One day, to begin our story, a surgeon discovers that an assistant has made a horrible mistake. Two men, a Mr. Brown and a Mr. Robinson, had been operated on for brain tumors, and brain extractions had been performed on both of them. At the end of the operations, however, the assistant inadvertently put Brown's brain in Robinson's head, and Robinson's brain in Brown's head. One of these men immediately dies, but the other, the one with Robinson's body and Brown's brain, eventually regains consciousness. (…) Over a period of time he is observed to display all of the personality traits, mannerisms, interests, likes and dislikes, and so on that had previously characterized Brown, and to act and talk in ways completely alien to the old Robinson (Shoemaker, 1964, p. 23 f.).

We are now asked whether this individual is Brown or Robinson and thereby also to reach a verdict on whether our diachronic persistence consists in mental or in bodily continuity. The variable that Shoemaker officially changed in his setup may be paraphrased as 'medicine is so advanced that brain transplantations are technologically possible'. While this would, of course, also result in some other factors being modified (people will, for example, have higher life expectancies), it may be reasonable to assume that, unlike in the case of teletransportation, which is located in a much more distant possible world, these alterations would not be far-reaching enough to significantly influence the conclusion. Consequently, the thought experiment's objectivity does not appear to be compromised. Of the many hypothetical situations that philosophers have devised to test their hypotheses about personal identity, whole-brain transplantations are certainly among the least demanding ones. A possible world in which this operation can be carried out is reasonably close.

Here, the problem is a different one. The setup presupposes that a specific relation obtains, in the actual world, between the brain and the rest of body: it is assumed that a particular organism does not exert a significant influence on mental features—and vice versa.[15] We are invited to decide between physical properties (remaining with the

[15] Bernard Williams (1970, p. 161) raised the worry that the new body might not be able to *express* the personality traits that it now houses. I do not think that this in itself presents a difficulty as there are other situations in which we accept that a particular individual is present despite even a total absence of motor output as, for example, locked-in syndrome shows. But what if the target of the cerebral transplantation was an organism of the opposite sex (Steinhart 2001, p. 21 n. 12)? In this case, the receiving body would subject the implanted brain to radically different endocrine influences.

body) and psychological properties (relocated with the brain), without considering the possibility that the latter may depend on the former. The standard reply would be that this should not matter precisely because we are dealing with a *hypothetical* situation, not with reality. One must look closely to see why this is not so: the counterfactual assumptions that this thought experiment makes are not supposed to extend to human physiology but to remain within the realm of technological advancements. If it is simply taken for granted that a brain would behave sufficiently alike in a different bodily environment, the reasoning becomes circular: in tacitly conjecturing that the body would not exert an influence on the brain that would be relevant to the person's psychological identity, one is begging the question against the advocates of biological views; for whether a person would persist when his or her brain was separated from the original organism and transferred to a different body is exactly what is at issue. Whether or not, or to which extent, bodies influence their brains and the mental processes to which the latter give rise is therefore highly relevant to the verdict that the thought experiment delivers if a *petitio principii* is to be avoided.

Since this question is a genuinely empirical one, engaging in purely conceptual speculation about this point is futile. One may well have an intuition as to what would happen if a brain was transplanted, but this intuition could easily be false. Brain and body are intimately connected via the nervous-, the endocrine-, and the vascular systems, through which they constantly exchange electrical impulses and chemical substances (Meier, 2020a, p. 20). It could therefore turn out that the interaction between the brain and the rest of the body was so peculiar to a certain organism that in the new environment the brain could not give rise to mental properties at all.[16] That one can *imagine* existing in Napoleon's body or that one can *coherently entertain* the possibility of being a brain in a vat does not mean that thought experiments based on such conjectures yield meaningful conclusions. The problem of being unfamiliar with the respective empirical findings is, as Kathleen Wilkes (1999, p. 19 f.) criticises,

> particularly pertinent to thought experiments concerning personal identity, precisely because most of the thought experimenters know little (and unfortunately care less) about biology and physiology (…), and relevant obstacles to the derivation of the conclusion (…) will be ignored.

To ensure that this is not just a feature of the specific thought experiment selected for discussion, I shall analyse another prominent example from the literature. Originally suggested by David Wiggins (1967, p. 53) and later developed by Derek Parfit, authors usually employ it to evaluate psychological criteria of personal identity under the condition of fission.

> My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my

[16] Recent findings seem to suggest that this is in fact not so (Meier, 2020b, Sect. 4.4). The point is, however, that one is not here dealing with an a priori proposition at which one can arrive without taking into account neurophysiological data. One may, of course, try to minimise the effects of such confounding variables by conducting one's thought experiment with the bodies of identical twins, as Jeff McMahan (2002, p. 20) does. This is a clever move, which, however, is not entirely free from empirical assumptions either: even identical twins are, strictly speaking, not phenotypically identical. And the question to which extent phenotypical differences affect mental characteristics is, once again, an empirical one.

> brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me (Parfit, 1987, p. 254 f.).

The brainstem houses the ascending reticular activating system, which regulates an individual's wakefulness level. When this neural network is destroyed, irreversible coma ensues (Hassler, 1971, p. 27 f.; Meier, 2020c, p. 100 f.; Moruzzi & Magoun, 1949, p. 471; Plum & Posner, 1980, p. 12). For Parfit's thought experiment to work, one must therefore make the empirical assumption that brainstems can be divided without rendering them functionless, so that each half can be transplanted together with the respective cerebral hemisphere. Parfit (1987, p. 255) acknowledges that 'it seems likely that it would never be possible to divide the lower brain, in a way that did not impair its functioning', but he contends that this did not matter as this impossibility would be 'merely technical'. In this case, one of the physiological background assumptions is thus made explicit, which is very helpful. What, however, does it mean for an impossibility to be merely technical and how does this bear on the strength of the conclusion derived?

One may distinguish several notions of impossibility. A case of *logical* impossibility would obtain if, for example, one and the same thing was both a person and not a person. By *metaphysical* (or conceptual) impossibility we mean what 'could not happen given our backing scientific knowledge: what our theories [do not] allow to be possible' (Wilkes, 1999, p. 18). Occurrences that are *physically* (or nomologically) impossible are not in accordance with the laws of nature. *Technical* impossibility, finally, is the weakest of these notions, or, in other words, the one located in the closest possible world. It denotes something that is logically sound and physically possible as well as in accordance with scientific concepts, but that has not been realised due to contingent reasons like a temporary lack of scientific progress. It is possible in principle.

Removing this obstacle in a thought experiment would consequently only necessitate the modification of the level of technological advancement as compared to the one prevalent in the actual world. It would not require that one tinker with any laws of nature. Prima facie, adjustments of this type should not pose a problem when the thought experiments in which they feature are designed carefully.[17] But does the procedure on which Parfit's thought experiment relies—the division of the brainstem—really fall under this category? Is it really only a contingent technical difficulty that prevents its realisation in the actual world, such as a lack of available surgical instruments?

Unlike the cerebrum, the brainstem is not a paired organ. The nuclei of the ascending reticular activating system, which serves both cerebral hemispheres simultaneously, are interrelated in a manner that precludes any attempt at slicing them in half without destroying this delicate structure (Laureys, 2005, p. 557). It is therefore not the case that if only we had at our disposal more sophisticated medical equipment, we could create two separately functioning ascending reticular activation systems out of one brainstem (Wilkes, 1999, p. 38 f.). While it may be logically possible to divide a whole brain and obtain two independently functioning slices, it is, for all we know, a

[17] In the previous sections, I have shown how even seemingly small alterations can impact on more variables than was originally intended, thereby rendering unintelligible the concepts against which the hypothesis is to be tested. But this need not be so in all cases.

physical and consequently a biological impossibility.[18] These are not merely technical hurdles as Parfit maintained.

Whether we are in this imagined procedure considering a philosophically useful situation or just something that we can somehow *conceive* depends on neuroanatomical and physiological properties of the brainstem—and thus on empirical facts that lie beyond what thought experiments can reliably establish or presuppose. As Adina Roskies (2016, p. 592) remarks: 'Insofar as philosophy aims to tell us about the world we live in, it is (or should be) as bound by fact as other disciplines'. Without taking into account biological facts, one is doing philosophy in the realm of science fiction, and it is highly doubtful that conclusions arrived at in this way are trustworthy guides to *our* persistence conditions. Not all questions can be answered from the armchair.

4 Conclusion

Most philosophers regard thought experimentation as the method of choice for illuminating the metaphysics of personal identity. In this paper, I have been investigating the strengths and weaknesses of this approach. I began with the question as to how it can be that one and the same hypothetical situation elicits contradictory intuitions and found that especially the more fantastical test scenarios that are often employed in an attempt to establish our persistence conditions fail to meet the quality standards of scientific experimental design. I argued that these standards also pertain to *hypothetical* experimentation when the imagined situations are so distant from the actual world that their premises become akin to empirical experiments in their own right. Simultaneously, however, the fancifulness of these hypothetical scenarios also precludes proceeding according to these principles.

Many thought experiments are not *objective* because in imagined worlds that are very different from the actual one, a multitude of uncontrolled variables—rather than only the purposefully manipulated independent variable—exert an influence on the outcome. It then becomes unclear from the modification of which parameters the observed change in the dependent variable originates.

Some fantastical thought experiments have yielded results that are interpersonally inconsistent. They are therefore not *reliable*. This is so because the more distant a possible world is from the actual one, the more non-*ceteris paribus* conditions demand specification. Although some situations are so bizarre that full accounts of the envisaged world's features would assume the length of books, descriptions of the imagined setups usually do not exceed a few paragraphs. It is therefore the subject carrying

[18] There is indeed a surgical procedure in which doctors sever the connections between the cerebral hemispheres (but leave the latter in situ). These so-called *commissurotomies* are sometimes cited as real-life counterparts of Parfit's thought experiment—for an overview of the different interpretations that have been proposed, see Tye (2003, pp. 111–113). *If* one regards the outcome of this operation as proof that consciousness can be divided, it would be best to work directly with the available clinical data instead of engaging in thought experimentation. Medically informed publications include Bayne (2008); Gillett (1986); Meier (2020b, pp. 116–119); Nagel (1985); Noonan (2003, p. 5 f.); Puccetti (1973, 1981); Schechter (2015); Wilkes (1999, pp. 132–167). What the procedure does unquestionably not show, however, is that the brainstem can be divided, since only pathways between the cerebral hemispheres are severed whereas the structures of the lower brain remain untouched.

out the respective thought experiment who must fill in these gaps, which makes the obtained result dependent on certain individual characteristics of the experimenter—an influence that is to be avoided in science at all cost.

Finally, the more distant a hypothetical scenario is, the less likely does it become that the conclusions drawn on the basis of the laws and concepts that obtain in the imagined setup are applicable to our world as it is. When asking questions about personal identity, we are normally inquiring about *our* persistence conditions. Conclusions derived from fantastical possible worlds therefore run the risk of not being *valid* when re-applied to the actual world.

From difficulties with *possible* worlds I moved on to unwarranted background assumptions about the *actual* world. Many popular thought experiments in the philosophical literature rely on unjustified presumptions about physiological details of the human body. While this is not a feature necessarily inherent to the method of thought experimentation since one could always take into account the available empirical facts, it appears that there has been little enthusiasm in the philosophical community for having the sheer endless options of imaginary setups constrained by anatomical or physiological limitations.[19] It may appear odd, writes Mark Johnston (2016, p. 92),

> to restrict our evidence base to the adventitious experiments of nature, when we could also avail ourselves of the full range of ingeniously designed thought experiments. Wouldn't that be like only considering the moves that have been made in actual chess games, rather than the full range of moves that *could have been* made?

Chess is only meaningful when it is played by the rules. Just like physical experiments, thought experiments should as strictly as possible adhere to the discussed standards of good scientific design: to objectivity, to reliability, and to validity—even if this means taking onboard much more empirical data than is customary in philosophy. Without rules, anything is possible; but the moves that one makes may become meaningless.

Thought experimentation remains a great tool for making scientific progress. In the natural sciences, this method has been pivotal to devising new theories and models. And in philosophy some of the most intriguing exchanges of arguments have sprung from the use of cleverly designed hypothetical situations. It is therefore not the hypothetical method as such that is questionable; nor even is it the hypothetical method applied to questions of personal identity. It is this method combined with *overly fantastical* scenarios. This is where the conclusions drawn become unreliable or even entirely inapplicable to the actual world.

Conducting thought experiments is one of the most important trademarks of philosophical thinking. No other discipline muses about individuals who can disconnect and reunite their cerebral hemispheres at will (Parfit, 1971, p. 6), exchange their brain states (Shoemaker & Swinburne, 1984, p. 108 f.), or have a dead tree in a swamp

[19] Philosophical works that rely mainly on empirical data instead of thought experimentation in deciding questions of personal identity include Campbell & McMahan (2010); Damasio (2010); McMahan (2002); Meier (2020b, 2020c); Nagel (1985); Puccetti (1969); Reid (2016); Savulescu & Persson (2016); Steinhart (2001); Strohminger & Nichols (2015). The most systematic approach is Kathleen Wilkes's (1999) monograph *Real People: Personal Identity without Thought Experiments*.

turn 'entirely by coincidence' into a physical replica of Donald Davidson (Davidson, 1987, p. 443). We need not abandon this fascinating method of reasoning; nor, however, should we overrate its power or underestimate its limitations.

Acknowledgements I would like to thank Katherine Hawley, Michael Wheeler, Christian Tewes, Thomas Fuchs, Peter Sullivan, Colin Johnston, Simon Prosser, and two anonymous reviewers of this journal. I am also grateful to Churchill College, Cambridge, and to the German National Academic Foundation for their support.

Funding This work was funded by Churchill College, Cambridge, and by the German National Academic Foundation.

Declarations

Conflict of interest All Authors declare that they have no conflict of interest.

References

Alexander, J. (2010). Is experimental philosophy philosophically significant? *Philosophical Psychology, 23*(3), 377–389. https://doi.org/10.1080/09515089.2010.490943

Bayne, T. (2008). The unity of consciousness and the split-brain syndrome. *The Journal of Philosophy, 105*(6), 277–300. https://doi.org/10.5840/jphil2008105638

Beck, S. (2014). Transplant thought-experiments: Two costly mistakes in discounting them. *South African Journal of Philosophy, 33*(2), 189–199. https://doi.org/10.1080/02580136.2014.923685

Bertram, G. W. (2012). Fiktionale Reflexionen von Begriffen: Zur Theorie philosophischer Gedankenexperimente. In G. W. Bertram (Ed.), *Philosophische Gedankenexperimente: Ein Lese- und Studienbuch* (pp. 24–72). Reclam.

Blatti, S. (2012). A new argument for animalism. *Analysis, 72*(4), 685–690. https://doi.org/10.1093/analys/ans102

Blatti, S. (2019). Animalism. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/animalism

Blatti, S., & Snowdon, P. F. (Eds.). (2016). *Animalism: New Essays on Persons, Animals, and Identity*. Oxford University Press.

Brown, J. R. (2004). Peeking into Plato's heaven. *Philosophy of Science, 71*(5), 1126–1138. https://doi.org/10.1086/425940

Brownlee, K., & Stemplowska, Z. (2017). Thought experiments. In A. Blau (Ed.), *Methods in Analytical Political Theory* (pp. 21–45). Cambridge University Press.

Buckwalter, W., & Turri, J. (2018). Moderate scientism in philosophy. In J. de Ridder, R. Peels, & R. van Woudenberg (Eds.), *Scientism: Prospects and Problems* (pp. 280–300). Oxford University Press.

Campbell, T., & McMahan, J. (2010). Animalism and the varieties of conjoined twinning. *Theoretical Medicine and Bioethics, 31*(4), 285–301. https://doi.org/10.1007/s11017-010-9150-0
Carruthers, P. (1995). *Introducing Persons: Theories and Arguments in the Philosophy of Mind*. Routledge.
Cohnitz, D. (2006). *Gedankenexperimente in der Philosophie*. Mentis.
Cole, D. (2020). The Chinese Room Argument. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/chinese-room/
Damasio, A. R. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon.
Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441–458. https://doi.org/10.2307/3131782
DeGrazia, D. (2005). *Human Identity and Bioethics*. Cambridge University Press.
Dennett, D. C. (1998). Where am I? In *Brainstorms: Philosophical Essays on Mind and Psychology* (pp. 310–323). MIT Press.
DePaul, M. R., & Ramsey, W. M. (Eds.). (1998). *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield.
Duhem, P. (1906). *La théorie physique: Son objet et sa structure*. Chevalier & Rivière.
Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik, 322*(10), 891–921. https://doi.org/10.1002/andp.19053221004
Galilei, G. (1974). *Two New Sciences: Including Centers of Gravity & Force of Percussion*. (S. Drake, Trans.). University of Wisconsin Press.
Gendler, T. S. (2000). *Thought Experiment: On the Power and Limits of Imaginary Cases*. Garland.
Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121
Gillett, G. (1986). Brain bisection and personal identity. *Mind, 95*(378), 224–229. https://doi.org/10.1093/mind/XCV.378.224
Green, M. B., & Wikler, D. (1980). Brain death and personal identity. *Philosophy & Public Affairs, 9*(2), 105–133.
Häggqvist, S. (1996). *Thought Experiments in Philosophy*. Almqvist & Wiksell.
Hannon, M. (2018). Intuitions, reflective judgments, and experimental philosophy. *Synthese, 195*(9), 4147–4168. https://doi.org/10.1007/s11229-017-1412-1
Hassler, R. (1971). Regulation der psychischen Aktivität. In R. Kurzrock (Ed.), *Hirnforschung und Psychiatrie: Stand und Zukunftsaspekte naturwissenschaftlicher Forschung in Deutschland* (pp. 27–35). Colloquium.
Higgins, A., & Dyschkant, A. (2014). Interdisciplinary collaboration in philosophy. *Metaphilosophy, 45*(3), 372–398. https://doi.org/10.1111/meta.12091
Ichikawa, J. J. (2013). Experimental philosophy and apriority. In A. Casullo & J. C. Thurow (Eds.), *The A Priori in Philosophy* (pp. 45–66). Oxford University Press.
van Inwagen, P. (1995). *Material Beings*. Cornell University Press.
Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly, 32*(127), 127–136. https://doi.org/10.2307/2960077
Johnston, M. (2016). Remnant persons: Animalism's undoing. In S. Blatti & P. F. Snowdon (Eds.), *Animalism: New Essays on Persons, Animals, and Identity* (pp. 89–127). Oxford University Press.
Knobe, J. (2007). Experimental philosophy. *Philosophy Compass, 2*(1), 81–92. https://doi.org/10.1111/j.1747-9991.2006.00050.x
Kolak, D. (1993). The metaphysics and metapsychology of personal identity: Why thought experiments matter in deciding who we are. *American Philosophical Quarterly, 30*(1), 39–50.
Kuhn, T. S. (1978). A function for thought experiments. In *The Essential Tension: Selected Studies in Scientific Tradition and Change* (pp. 240–265). University of Chicago Press.
Kühne, U. (2005). *Die Methode des Gedankenexperiments*. Suhrkamp.
Laureys, S. (2005). The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences, 9*(12), 556–559. https://doi.org/10.1016/j.tics.2005.10.010
Lewis, D. (1976). Survival and identity. In A. O. Rorty (Ed.), *The Identities of Persons* (pp. 17–40). University of California Press.
Liao, S. M. (2008). A defense of intuitions. *Philosophical Studies, 140*(2), 247–262. https://doi.org/10.1007/s11098-007-9140-x
Lienert, G. A., & Raatz, U. (1994). *Testaufbau und Testanalyse* (5th ed.). Psychologie Verlagsunion.
Locke, J. (2008). *An Essay Concerning Human Understanding*. P. Phemister (Ed.). Oxford University Press.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*(1), 128–159. https://doi.org/10.1111/j.1475-4975.2007.00160.x
Mach, E. (1926). *Erkenntnis und Irrtum: Skizzen zur Psychologie der Forschung* (5th ed.). Barth.
Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition, 92*(3), B1–B12. https://doi.org/10.1016/j.cognition.2003.10.003
McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.
Meier, L. J. (2020a). The demise of brain death. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz045
Meier, L. J. (2020b). *Brain Death: What We Are and When We Die*. University of St Andrews. https://doi.org/10.17630/sta/17
Meier, L. J. (2020c). Are the irreversibly comatose still here? The destruction of brains and the persistence of persons. *Journal of Medical Ethics, 46*(2), 99–103. https://doi.org/10.1136/medethics-2019-105618
Merricks, T. (2001). *Objects and Persons*. Clarendon Press.
Moruzzi, G., & Magoun, H. W. (1949). Brain stem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology, 1*(4), 455–473. https://doi.org/10.1016/0013-4694(49)90219-9
Nagel, T. (1985). Brain bisection and the unity of consciousness. In *Mortal Questions* (pp. 147–164). Cambridge University Press.
Nelson, A. A. (1980). Research design: Measurement, reliability, and validity. *American Journal of Hospital Pharmacy, 37*(6), 851–857. https://doi.org/10.1093/ajhp/37.6.851
Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology, 23*(3), 293–312. https://doi.org/10.1080/09515089.2010.490939
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x
Nida-Rümelin, M., & O Conaill, D. (2019). Qualia: The Knowledge Argument. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/qualia-knowledge/
Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310. https://doi.org/10.1037/0033-295X.108.2.291
Noonan, H. W. (2003). *Personal Identity* (2nd ed.). Routledge.
Norton, J. D. (1996). Are thought experiments just what you thought? *Canadian Journal of Philosophy, 26*(3), 333–366. https://doi.org/10.1080/00455091.1996.10717457
Olson, E. T. (1997). Was I ever a fetus? *Philosophy and Phenomenological Research, 57*(1), 95–110. https://doi.org/10.2307/2953779
Olson, E. T. (1999). *The Human Animal: Personal Identity without Psychology*. Oxford University Press.
Ørsted, H. C. (1998). *Selected Scientific Works of Hans Christian Ørsted*. K. Jelved, A. D. Jackson, & O. Knudsen (Eds.). Princeton University Press.
Parfit, D. (1971). Personal identity. *The Philosophical Review, 80*(1), 3–27. https://doi.org/10.2307/2184309
Parfit, D. (1987). *Reasons and Persons*. Clarendon Press.
Parfit, D. (1995). The unimportance of identity. In H. Harris (Ed.), *Identity* (pp. 13–45). Oxford University Press.
Perry, J. (1972). Can the self divide? *The Journal of Philosophy, 69*(16), 463–488. https://doi.org/10.2307/2025324
Plato. (1997). *Complete Works*. J. M. Cooper & D. S. Hutchinson (Eds). Hackett.
Plum, F., & Posner, J. B. (1980). *The Diagnosis of Stupor and Coma* (3rd ed.). Davis.
Puccetti, R. (1969). Brain transplantation and personal identity. *Analysis, 29*(3), 65–77. https://doi.org/10.1093/analys/29.3.65
Puccetti, R. (1973). Brain bisection and personal identity. *The British Journal for the Philosophy of Science, 24*, 339–355.
Puccetti, R. (1981). The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences, 4*(1), 93–99. https://doi.org/10.1017/S0140525X00007755
Putnam, H. (2000). Brains in a vat. In S. Bernecker & F. I. Dretske (Eds.), *Knowledge: Readings in Contemporary Epistemology* (pp. 1–21). Oxford University Press.
Reid, M. D. (2016). A case in which two persons exist in one animal. In S. Blatti & P. F. Snowdon (Eds.), *Animalism: New Essays on Persons, Animals, and Identity* (pp. 253–265). Oxford University Press.
Rescher, N. (2005). *What If? Thought Experimentation in Philosophy*. Transaction.

Roskies, A. L. (2016). Neuroscience. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.), *The Oxford Handbook of Philosophical Methodology* (pp. 587–606). Oxford University Press.
Savulescu, J., & Persson, I. (2016). Conjoined twins: Philosophical problems and ethical challenges. *The Journal of Medicine and Philosophy, 41*(1), 41–55. https://doi.org/10.1093/jmp/jhv037
Schechter, E. (2015). The subject in neuropsychology: Individuating minds in the split-brain case. *Mind & Language, 30*(5), 501–525. https://doi.org/10.1111/mila.12088
Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Die Naturwissenschaften, 23*(48), 807–812. https://doi.org/10.1007/BF01491891
Schwartz-Shea, P., & Yanow, D. (2012). *Interpretive Research Design: Concepts and Processes*. Routledge.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756
Shoemaker, S. (1964). *Self-Knowledge and Self-Identity*. Cornell University Press.
Shoemaker, S. (1970). Persons and their pasts. *American Philosophical Quarterly, 7*(4), 269–285.
Shoemaker, S. (1999). Self, body, and coincidence. *Aristotelian Society Supplementary, 73*(1), 287–306. https://doi.org/10.1111/1467-8349.00059
Shoemaker, S. (2008). Persons, animals, and identity. *Synthese, 162*(3), 313–324. https://doi.org/10.1007/s11229-007-9253-y
Shoemaker, S., & Swinburne, R. (1984). *Personal Identity*. Blackwell.
Sider, T. (2001). Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives, 15*, 189–209. https://doi.org/10.1111/0029-4624.35.s15.10
Snowdon, P. F. (1990). Persons, animals, and ourselves. In C. Gill (Ed.), *The Person and the Human Mind: Issues in Ancient and Modern Philosophy* (pp. 83–107). Clarendon Press.
Sorensen, R. A. (1992). *Thought Experiments*. Oxford University Press.
Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies, 132*(1), 99–107. https://doi.org/10.1007/s11098-006-9050-3
Steinhart, E. (2001). Persons versus brains: Biological intelligence in human organisms. *Biology and Philosophy, 16*(1), 3–27. https://doi.org/10.1023/A:1006718411266
Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26*(9), 1469–1479. https://doi.org/10.1177/0956797615592381
Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research, 76*(1), 138–155. https://doi.org/10.1111/j.1933-1592.2007.00118.x
Talbot, B. (2013). Reforming intuition pumps: When are the old ways the best? *Philosophical Studies, 165*(2), 315–334. https://doi.org/10.1007/s11098-012-9949-9
Tetens, H. (2016). Reproducibility, objectivity, invariance. In H. Atmanspacher & S. Maasen (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects* (pp. 13–20). Wiley-Blackwell.
Thompson, E., & Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical Topics, 39*(1), 163–180. https://doi.org/10.5840/philtopics201139119
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217. https://doi.org/10.5840/monist197659224
Tobia, K. P. (2015). Personal identity and the Phineas Gage effect. *Analysis, 75*(3), 396–405. https://doi.org/10.1093/analys/anv041
Tye, M. (2003). *Consciousness and Persons: Unity and Identity*. MIT Press.
Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*(6), 476–491.
Unger, P. (1990). *Identity, Consciousness and Value*. Oxford University Press.
Weinberg, J. M., & Crowley, S. (2009). The X-Phi(les): Unusual insights into the nature of inquiry. *Studies in History and Philosophy of Science, 40*(2), 227–232. https://doi.org/10.1016/j.shpsa.2009.03.016
Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29*(1–2), 429–460. https://doi.org/10.5840/philtopics2001291/217
Wiggins, D. (1967). *Identity and Spatio-Temporal Continuity*. Blackwell.
Wiggins, D. (1980). *Sameness and Substance*. Blackwell.
Wiggins, D. (2003). *Sameness and Substance Renewed*. Cambridge University Press.
Wilkes, K. V. (1999). *Real People: Personal Identity without Thought Experiments*. Clarendon Press.
Williams, B. (1970). The self and the future. *The Philosophical Review, 79*(2), 161–180. https://doi.org/10.2307/2183946
Williams, B. (1973). *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge University Press.

Wittgenstein, L. (1967). *Zettel*. (G. E. M. Anscombe, Trans., G. E. M. Anscombe & G. H. von Wright, Eds.). University of California Press.