5 Thought experiments, concepts, and conceptions

Daniele Sgaravatti

1 Introduction

The reader will most likely be familiar with the presentation of a thought experiment in a philosophical text. A scenario, or hypothetical case, is described. Usually, a judgment is given about what would be true in the scenario (sometimes called the "intuition" or the "intuitive" judgment). Some conclusions are drawn from that judgment, together with the (sometimes implicit) judgment that the scenario is possible.[1] Gettier's cases of justified true belief falling short of knowledge, Kripke's counterexamples to descriptivist theories of the meaning of proper names, Jackson's case of the colour-deprived neuroscientist Mary, Foot's morally troubling trolley cases, and so on, are paradigmatic examples. Several questions can be raised about the judgments philosophers give on thought experiments. Experimental philosophers claim to have provided evidence that these judgments are unreliable. My focus here will not be on that challenge. I do not think the challenge is empirically sound, but nothing I say here depends on that. There is a general question that should be addressed, it seems, before addressing the reliability question, or at least independently of it. What are judgments about thought experiments based on? Some philosophers think that the answer to this question will require appeal to the notion of intuition as a *sui generis* mental or epistemic kind, or to the notion of (epistemic) analyticity. On the former kind of view, judgments about thought experiments are based on a sort of intellectual perception; on the latter, they are justified merely in virtue of the subject's understanding of the relevant words or concepts. I will assume here that there are serious problems for both strategies, and that an answer that does not require appeal to these notions is therefore to be preferred, at least prima facie. The view I will defend here avoids those kinds of heavyweight theoretical commitments. In giving judgments about thought experiments, on my view, we just use our basic ability to apply concepts, the same ability we employ in giving judgments about

1 It is not uncontroversial that the judgment about a case can be separated as I am doing from the modal judgment that the case is possible (see, e.g., Malmgren 2011). I will not have space here to defend that assumption. I suggest however that reflection on cases in which we can find an actual instantiation of the hypothetical scenario strongly supports it.

actual situations. A similar claim is made by Williamson where he writes: "we assent to [the Gettier verdict] on the basis of an offline application of our ability to classify people around us as knowing various truths or as ignorant of them, and as having various other epistemologically relevant properties" (Williamson 2007, 188).

In the next section I will look at some broad theoretical characterizations of thought experiments, and I will introduce my proposal, which is that judgments about thought experiments constitute knowledge when they are correct because of a competence in applying the relevant concepts. In the following section I will concentrate on what I take to be the main challenge for my sort of view, namely that of giving a characterization of competence in applying concepts which (a) does not resort to the notions of 'intuition' or 'analyticity' and (b) is not vacuous. In the third section (pp. 000–000), I will look at psychological theories of concepts, and I will argue that they can help my proposal when they are interpreted as theories of our capacity to apply concepts, as they should be for independent reasons. In the last section, I will further clarify my view by looking at an objection and by comparing it with a similar one developed in Papineau (2009).

2 A simple view of thought experiments, and a challenge

The view that judgments about thought experiments only require our ordinary ability to apply concepts is supported by reflection on some attempts to define the notion of a thought experiment. What is a thought experiment, exactly? Brown and Fehige, in the related entry in the *Stanford Encyclopaedia of Philosophy*, write that "[t]hought experiments are devices of the imagination used to investigate the nature of things" (Brown and Fehige 2011). Sorensen defines a thought experiment as "an experiment … that purports to achieve its aim without the benefit of execution" (Sorensen 1992, 205). He refers back to the following definition of experiment:

> An experiment is a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others. For the sake of simplicity, most experiments are designed around two variables. The one you directly manipulate is called the independent variable and the one you try to affect indirectly through these manipulations is the dependent variable.
>
> (Sorensen 1992, 186)

It is noteworthy just how liberal these definitions are. I might imagine that I dropped my cup of coffee, and judge that in that case the cup would have fallen to the ground and spilled its content on the floor. Or I could imagine that there is a law which prohibits comparing philosophical thought experiments to any kind of reasoning occurring outside philosophy, and judge that in that case I would be breaching the law right now. Are these thought experiments? Williamson (2007) suggests that there is no essential difference between thought experiments in

philosophy and ordinary counterfactual judgments, e.g. 'if I dropped my cup of coffee, it would fall on the ground and spill its content'. The latter sentence fits, although somewhat loosely, Brown's definition. I might be using a device of the imagination to investigate the nature of my cup of coffee, or perhaps the nature of physical laws in our universe. Similarly, the reasoning described fits Sorensen's definition; I am affecting (in imagination) the independent variable of my holding the cup, and observing an effect on the dependent variables constituted by the cup and its content.

The foregoing considerations suggest that many difficulties in the epistemology of thought experiments may derive from our tendency to think that, because the scenarios most prominent in philosophy are somewhat out of the ordinary, we must be employing extraordinary means in dealing with them. In so doing, we overlook certain similarities between the thought experiments used in philosophy and more common acts of thought. Paradigm cases are sometimes misleading, as we know. To use an analogy, suppose we first identified planets as some small lights (certain 'stars', as we might have said) which were moving in a somewhat irregular way in the night sky, knowing next to nothing else about the nature of these celestial bodies. In that situation, the suggestion that the ground on which we were standing was itself a planet might have seemed implausible, or even absurd; it did not fit our way of individuating planets in the paradigmatic cases. We could be making a similar mistake in thinking that thought experiments in philosophy are radically different from more ordinary specimens of hypothetical reasoning; we might be failing to look at the epistemic ground on which we stand.

However, even if one is sympathetic to the idea that it is our ordinary ability to apply concepts that guides out judgments about thought experiments, one might still worry that it lacks in explanatory power, or that it amounts to little more than a metaphor. That we have the ability to apply concepts, in general, should not be contentious. But one might complain that just talking about our ability to employ concepts in order to explain how we form judgments is no more explanatory than talking about a certain substance's *virtus dormitiva* in order to explain how that substance causes people to fall asleep. It will be useful to distinguish different ways to understand the challenge. In so doing, I will also start to make the view I am endorsing more precise.

On one hand, one might be asking for an epistemological theory that tells us when judgments in general are justified, or constitute knowledge. Developing such a theory is beyond the scope of the present work. But I also think it is not the point of the question. One can fairly easily find a plausible epistemological view that allows for our hypothetical judgments to constitute knowledge, but this does not in itself provide the theory with explanatory power. I will adopt, for definiteness, a view similar to the one presented in Sosa (2007), on which knowledge is *apt* belief, and apt belief is defined as a belief that is correct because it manifests an epistemic virtue or competence, which is, roughly, a capacity to discriminate true contents from false ones. Adopting Sosa's general framework, the (simple) view I wish to defend is the following:

> Simple Competence View (SCV): A judgment about a thought experiment constitutes knowledge if and only if it is correct because it manifests a competence in applying the relevant concepts.

Two important points of clarification. Firstly, one could doubt the correctness of the general epistemological view, as it often happens, because of putative counterexamples to the sufficiency of the condition. Even so, one could take the right-hand side of the biconditional as stating an interesting necessary condition on knowledge via thought experiments, and the rest of this paper as an attempt at explaining how the condition can be met. Secondly, I wish at this point to stress that it is not part of the view that possessing a concept requires having a competence in applying it. The competence requires a certain degree of reliability, which is not guaranteed by simple possession of the concept.[2]

SCV differs from Sosa's own view, which is that judgments about philosophical thought experiments, or at least some of them, are of a special kind, because first, they are confined to modally strong contents (necessarily true or necessarily false) and, second, they are explained by a competence that does not rely on perception, memory, introspection, testimony or inference. Call that the Rationalist Competence View (RCV). I believe the additional requirements of RCV have to be dropped, for reasons that I will explain (at least in part) in the last section. In my view, the judgments on philosophical thought experiments might not be typically based directly on perception, introspection, testimony, episodic memory, or explicit inference, but the competence itself is epistemically dependent on some of these sources, at least in standard cases.

However, here I am interested in an objection that applies (although perhaps not with the same strength) both to RCV and to SCV. The objection spells out the worry about explanatory power mentioned above, and it is put forward by Boghossian (2009). Boghossian writes that "it can look as though we have invoked a mystery to explain a mystery".[3] We have been told nothing about how the relevant competence works. We asked how we obtain knowledge through thought experiments and, the objection would go, the only answer we had is that we possess a power to obtain knowledge from thought experiments. The appeal to "competence in applying a concept" is tantamount to appeal to *virtus cognitiva*—so I will refer henceforth to the problem as 'the *virtus cognitiva* objection'. Part of the problem, according to Boghossian, is that Sosa's competence is supposed to work based only on the understanding of the proposition. I agree that this makes the problem more severe for RCV than for SCV, but it seems clear that the objection could be directed to both. The proposal to be developed in the next section should clearly answer the *virtus cognitiva* objection. Before moving to my proposal, I will consider Sosa's own reply, and explain why I take it to be not entirely

2 This is why I am not appealing, even implicitly, to the notion of analytic judgment.

3 Boghossian 2009, 116. Malmgren (2011) develops a similar sort of objection against Williamson's view of thought experiments.

satisfactory. This will also constitute one more effort in clarifying the challenge. In reply to Boghossian's point, Sosa writes:

> as I have argued elsewhere, we can appeal to a sort of competence in epistemology even when we have only limited understanding of its modus operandi. This in fact applies not only to rational intuition but also to introspection and even to perception and memory. People could surely know how they knew things even before we gained our vastly improved understanding of how perception and memory actually work. Of course, our knowledge through these various sources will be enhanced with our improved understanding of their nature and operation.
>
> (Sosa 2009, 140)

Sosa is certainly correct that people with little or no understanding of the workings of, say, visual perception, can acquire knowledge through its use, and even come to know that they have acquired knowledge and, under a certain description, how they acquired it ('how do you know that? I saw it'), thereby possessing, in Sosa's terms, *reflective* knowledge of the propositions in question. In Sosa's view, reflective knowledge of a proposition requires at least aptly believing that one aptly believes the proposition. However, this is not yet to say that we can appeal to that competence in epistemological theorizing. We can of course use the general notion of competence; but it seems hard to appeal to a specific competence without being able to explain at least to some degree how it works. Saying 'I saw it' can be a perfectly good explanation of how one came to know something in some contexts, but it is not a satisfying epistemology. Something more is required.

Sosa also claims that reflective knowledge comes in degrees, and the higher degrees "may of course involve scientific and even philosophical perspectives that enable defence of one's first-order belief as apt" (Sosa 2009, 141). This is probably what he is hinting at when he says that improved understanding of the nature of a competence will "enhance" our knowledge. However, this enhancement is precisely the object, or one of the main objects, of epistemological theorizing; and our epistemology fails to provide it if it fails to integrate our legitimate knowledge claims into a broader scientific picture. Let us consider again the parallel with perception. Philosophers such as Aristotle, Descartes and Hume were certainly interested in the psychology of perception. Of course, psychology was not at the time an independent discipline. But their interest in this area does not seem, even today, disconnected from their philosophical views. They were interested in providing a coherent and comprehensive world picture, and in particular an account of the place of human minds in the world. Suppose for example that a philosopher defends a radical form of Cartesian dualism about the mind. Could she then defend an epistemological theory of perception that only says that a perceptual belief is justified when the subject's perceptual capacities are reliable? Clearly, this would be an incomplete picture. We would not know *how* the immaterial mind gets in touch with the world in a reliable way. If a dualist is without an answer to this question, her epistemology is defective, even if it does not, strictly speaking,

entail that we do not gain knowledge by perception. In a way, I am attempting to answer a question of the same kind (although hopefully not a similarly desperate one) about our capacity to apply concepts. The question is not purely epistemological, but it is not purely a matter of psychology or cognitive science either. It is a problem at the border between these disciplines, or, one might say, a problem at the interface of philosophy and cognitive science.

3 The psychology of (the application of) concepts

In this section I will look at some of what psychology and cognitive science tell us about the way we apply concepts. In particular, I will look at various theories of concepts, and I will argue that this empirical literature provides theories that, although not originally intended this way, can show what an ability to apply a concept is (without appealing to the notions of 'intuition' or 'analyticity'). A few notes of caution are in place. First, there is in this area a vast amount of disagreement. We cannot just consult *the* correct theory of concepts and check what consequences it has for our philosophical concerns. My choice will be somewhat arbitrary.[4] Secondly, sometimes the philosophical interpretation of the theory is equally controversial. Thirdly, I will at first present the theories as if they were incompatible, but many psychologists nowadays prefer a pluralist approach, on which different accounts will hold for different concepts and even for the same concept on different occasions.[5] My way of reinterpreting such theories actually makes the pluralistic approach more plausible. Finally, what follows will necessarily be a very rough account of the theories I consider and of their strengths and defects. We could say that we are looking at toy versions of the theories (and toy versions of the objections). But this should still be sufficient for my present purposes. I will proceed by first sketching two different theories of concepts, then I will discuss some objections, and finally I will argue that, in the light of those objections, the theories I focus on are implausible as theories of concepts, but they might work as theories of our ability to apply concepts.

The first theory, or maybe better family of theories, of concepts I will look at is *prototype* theories. This kind of approach emerged in the 1970s due especially to the work of Eleanor Rosch (Rosch 1973, 1978; Rosch and Mervis 1975), as a reaction to the classical theory of concepts, according to which a concept is characterized by a set of necessary and sufficient conditions of applications. The

4 Other theories that I might have considered in this connection, going from less to more fashionable, are: the so-called *classical* theory of concepts, according to which concepts consist of a set of necessary and sufficient conditions; the *exemplar* theory of concepts, according to which concepts are, roughly, the storage of a number of exemplars, where each exemplar is identified by a set of properties it possesses; the *situated* or *embodied* theory of concepts, according to which concepts are, roughly, collections of action-oriented abilities. Although some of the details would have to vary (in particular, objections to the classical theory are quite different from objections to the other theories; see Laurence and Margolis 1999b), the conclusion that I am going to reach later about prototype theory and theory-theory could be defended about all of the foregoing theories.

5 See Machery 2009 for an extended treatment on this issue, but see also Rey 2009.

classical theory had come under attack from a variety of fronts; most notably it was thought to be incompatible with the Quinean criticism of the analytic–synthetic distinction. However, the philosophical inspiration for prototype theories was Wittgenstein's notion of a family resemblance concept. The core idea of the prototype theory is that a concept is structured as a set of features that the objects falling under the concept *tend* to have.[6] Crucially, none of the features is necessary. The concept can thus be thought of as structured around a prototype, a model of an object possessing all, or as many as possible, of the features. Anything resembling the prototype to a sufficient degree, i.e. having a sufficient number of features, will belong to the category. Membership comes in degrees; the more features something has, the more it belongs to the category. There are various ways of interpreting the latter claim, and they make serious theoretical differences, but the basic thought is that some things are better, or clearer, exemplars of the category than others. To illustrate, consider the concept BIRD. It will consist of a list of features including, let us suppose, "has a beak, flies, has wings, has plumage, lays eggs, build nests, sings, is small (compared to a human)". On this view, a robin is as close to the prototype as possible, having all the features. A chicken is less prototypical, but still a good case. An ostrich is quite far, and a penguin is probably at the far end.

The theory predicted a series of cognitive effects, which we will call *typicality* effects, that were found to obtain (see, e.g., Rosch and Mervis 1975; Laurence and Margolis 1999a). The main typicality effects are the following: (1) Graded membership: people are willing to classify objects in various categories as better or worse exemplars of a category, often with a blurred boundary between very bad cases and things outside the category. This ranking correlates with the other typicality effects. (2) Retrieval: if asked to name a member of the category, the subjects will name a prototypical member; and if they have to list a number of members, they will do so in order of typicality. (3) Speed and accuracy of categorization: if asked to decide whether an object belongs to the category, the subjects will reach a faster and more often accurate verdict in the case of more prototypical members, and they will give a slower and more often mistaken verdict for less typical members. (4) Correlation with features: moreover, the effects in (1), (2) and (3) can be predicted with an appropriate choice of features possessed by the prototypical members.

I will discuss problems for this theory mostly after I have presented the second kind of theory we are going to look at, since they affect both views. There is however an exception, an interesting problem which is specific to prototype theories. Armstrong *et al.* (1983) conducted experiments which showed typicality effects for concepts for which they thought the prototype theory seemed not adequate; in particular they tested the concepts ODD NUMBER, EVEN NUMBER, FEMALE, PLANE GEOMETRICAL FIGURE. For example they found that the subjects were

6 Later versions of the theory gave statistical models which specify numerically the "weight" of the feature in determining membership in the category, which made it possible to produce precise empirical predictions. I will concentrate on the simple version of the theory.

willing to classify 34 as a worse exemplar of an even number than 8. But, they thought, it is implausible that the concept EVEN NUMBER has a prototype structure. It seems that something is an even number if and only if it is a number and it is even, and neither condition admits of degrees. Both defenders and opponents of the prototype theory took the problem very seriously, recognizing it as a major difficulty for the view. However, Rosch takes a different line, and sees it as a "strange twist of logic" (Rosch 1999, 66) that the fact that the theory applies to those concepts as well was used as an argument against it. I think, and here I am anticipating my main conclusion about this debate, that Rosch's reply only makes sense if we understand prototype theory not as a theory of the content of the concepts, but rather as a theory of the way we apply concepts. Surely 34 is an even number just as much as 8 is, unless we subscribe to a strange form of anti-realism about even numbers; but the way we apply the concept EVEN NUMBER could well depend on taking some number as prototypically even.

The other kind of theories that I want to consider are *theory-theories* of concepts, i.e. the kind of theories according to which concepts are theories (in a sense to be specified). In a paper related to philosophical methodology, Cummins (1998) describes this kind of view as follows: "the majority view, I think, is that concepts are theories, either explicit, as in the case of technical scientific or legal concepts, or tacit, as in the case of 'ordinary' concepts." And he goes on to endorse this kind of view (combined with an element of prototype theory):

> My own view, for what it is worth, is that my concept of an elevator is just everything I know about elevators, different bits of which are activated or accessed on different occasions, depending on cues and previous activations, plus some quick and dirty procedures that account for prototype effects.
>
> (Cummins 1998, 121)

There are two important clarifications which have to be added to what Cummins says. Firstly, one could hold a theory-theory view on which the theory constituting the concept is innate, for at least some concepts. But the most common version of this kind of theory holds, quite reasonably, that most concepts are acquired, and somehow vary with our knowledge of the world. It is of course a matter of great interest how much of our conceptual apparatus is innate, but I am not presupposing any answer to that question. Secondly, there is much controversy about the nature of conceptual change. In an extreme form of the view, conceptual change is radical; as our concepts change, the new theories are incommensurable with the old ones, in the sense in which according to (a reading of) Kuhn scientific theories are incommensurable with old ones after a shift in the dominant paradigm (on both issues, see Carey 1991 and 2009).

Margolis and Laurence (1999b) cite as strengths of the theory-theory the ability to give a realistic account of categorization judgments and cognitive development. The former is for example found in the possibility of a theory-theory to account for the tendency of adults and, interestingly, children, to take an essentialist view of natural kinds. Contrary to what the prototype theory would plausibly predict,

children as well as adults think that external features of a member of a natural kind, such as dogs, are comparatively irrelevant to whether it is a member of the category, and what is relevant is some internal, essential property (Gelman and Wellman 1991).

It is now time to look at some objections. There are at least four problems discussed in the literature (Rey 1983; Fodor 1998 and 2008; Laurence and Margolis 1999b) that can be applied to both theories. They are (1) the problem of shareability of concepts, (2) the problem of ignorance, (3) the problem of compositionality and (4) the explanatory regress problem. I'll give a very brief sketch of each. Of course the dialectic could be pursued further, with different replies or modifications of the theories for each objection, but I will explain later why, problems of space aside, we do not need to do this.

The problem of shareability depends on the fact that prototypes as well as theories associated with a concept (or a word) vary both across subjects and for a single subject over time. However, if this is so, my concept BIRD is not the same as, say, the ornithologist's concept, and the ornithologist's concept is not the same as it was when she started her studies. I picked a rather extreme case for vividness of illustration; prototypes will vary across non-experts as well (see Barsalou 1987); and clearly theories vary across individuals and times, at least if we understand them as Cummins does. This means, it seems, that our beliefs and the expert's cannot be in contradiction, which is an unwelcome result, for surely we take the experts to be expert on the very subject matter on which we are not. I must confess it is mysterious to me how it came to be, in cognitive science, that a theory which makes it impossible for people to believe the negation of what they used to believe (since abandoning the previous belief requires a change in the concepts involved) is supposed to have a strength in its capacity to account for cognitive development. The problem is particularly severe for the radical version of the theory on which the new concept will often be a theory which is incommensurable with respect to the old one.

The problem of ignorance is substantially a form of the objection which was posed by Kripke, Putnam and others against descriptivist theories of names and natural kind terms. We can be radically wrong in our prototypes and theories, and still manage to refer to the relevant objects. If the theories in question wish to give some explanation of the connection between the way they describe the concept and the concept's reference, this seems to be a serious problem; of course they could just not wish to tell a story about it, but then they seem in that case to fail to meet an explanatory burden.

The problem of compositionality has been discussed especially in relation to prototype theories. A complex concept, such as PET FISH, will not have of course the union of the prototypical features of the component concepts (a prototypical fish lives in the sea, or in a river or lake; a prototypical pet can be cuddled). But neither is it simply the intersection; the prototypical pet fish lives in a bowl, but that feature is nowhere to be found for either pets or fishes. So it seems that, on the prototype theory, there just is no way in which the structure of a complex concept can be a function of the structure of the simple concepts on the prototype

theory. This is a huge problem, for it is clear that we can understand a number of new complex concepts far too vast to make it plausible that we are just learning a new independent entry each time. I know of little discussion of this particular problem as applied to theory-theories. But the reason does not seem to be that there is some straightforward account of how two theories compose a complex theory; on the contrary, it seems entirely obvious that there is no mechanical procedure for combining two theories.

The last problem we have to consider is the one I am calling "the explanatory regress problem". This is the only objection which I am considering that is not, as far as I know, widely discussed in the literature already—at least not in this form. Let us ask ourselves, with respect to the prototype theory: how is a feature represented? If a feature is just a property of the members, we would expect the subject to have a concept that applies to that property; otherwise the subject could hardly use the feature to compute the typicality of a member. But, if the concept of the feature is a list of features, and each of those is in a turn a concept, we can easily see that there would be an exponential regress. So the prototype theorist must either maintain that some concepts are not prototypical or that the subjects think of features in some way other than by having concepts. Unsurprisingly, the latter is the alternative prototype theorists prefer. Some features could be perceptual traits, recovered as images; but it would seem a naïve form of empiricism to suppose that a concept can be composed of perceptual features *only*. So there must be a way in which information about features is processed without being conceptualized; but it seems that we lack an explanation of how this is done. The problem is even more dramatic for theory-theories. A theory, we would normally think, is made of various integrated beliefs, and beliefs are made of concepts. If each concept is a theory, we seem again to be stuck in a regress. On the other hand, if we think of the information in the concept as somehow non-conceptually organized, talk of "theory" now seems purely metaphorical.

This completes my short survey of the two theories of concepts and of their problems. The problems considered seem to be extremely serious. Of course there are other theories of concepts in psychology, some related to these ones, some rather different; the two theories I have discussed here were not chosen because they are particularly plausible. In fact, I must now confess, they were chosen because they are, in my opinion, particularly *im*plausible, as theories *of concepts*. My suggestion is that we should think of them as theories of the ability to apply concepts.[7] Interpreted this way, the theories retain all the advantages they had, including the empirical evidence in their favour, for the phenomena they are meant to explain all involve the application of concepts. On the other hand, all the problems mentioned disappear once the theories are reinterpreted this way. Abilities to apply concepts are not supposed to be shared, and can vary (in fact, hopefully they improve) across cognitive development. They are not supposed to

7 A similar conclusion is reached by Rey (1983) about prototype theories. Rey's considerations clearly apply, *mutatis mutandis*, to theory-theories.

determine the reference of the concept. They are not supposed to combine according to syntactic rules. Moreover, they can consist of associations, either in the form of a prototype or of a tacit or explicit theory, between concepts. Finally, we can see how the two theories, as theories of the ability to apply concepts, are not incompatible but rather complementary. We may use prototypes in applying certain concepts and tacit theories in applying others. We also may use sometimes prototypes and sometimes tacit theories in applying a single concept, depending on the context; for example, in perceptual recognition we might use mainly prototypes, and in inferential processes mainly tacit theories. We may call the combination of a prototype and a theory (and any other psychological structure used in the application of the concept) a 'conception'.[8] Of course, not every conception produces justified beliefs; one can have a very misguided conception of an object or kind of object, thereby being very unreliable in applying the concept. On the other hand, clearly a conception can, in some cases, be sufficiently accurate to be a reliable guide to whether the concept applies. There will be still further conditions that the conception will have to satisfy to count as a competence, which will be discussed in the next section.

Given how easy it turns out to be to understand these theories as theories of conceptions, it might be thought that what I am suggesting is really superfluous, because the theories are already what I am suggesting they should become: psychologists really mean 'ability to apply a concept', or 'conception' when they say 'concept', and I am somehow reading them uncharitably. Although I think this stance would have some advantages, ultimately I do not endorse this interpretation of the content of the psychological theories I mentioned. But this is not at all crucial for my purposes. If theories of the ability to apply a concept were already there to be found, this would make my reply to the *virtus cognitiva* objection even easier.

However, I think it is worth pausing on this point. It is worth doing so both because it helps to clarify my stance, and because it is surely worth clarifying some confusions surrounding the concept CONCEPT. One could think that the word 'concept' is really ambiguous, and philosophers and psychologists are often talking past each other. Here are two characterizations of concepts, the first offered by a philosopher and the second by a (philosophically inclined) cognitive scientist:

> Concepts are sub-components of thought contents. Such contents type propositional mental events and abilities that may be common to different thinkers or constant in one thinker over time. Having a concept is just being able to think thoughts that contain the concept. … In being components of thought contents, and ways of thinking, concepts are representational or intentional (I make no distinction here). They need not apply to actual objects, but their function is such that they purport to apply; they have intentional or referential functions.
>
> (Burge 1993, 309–310)

8 I take the term, and the notion, from Millikan 2000. See also Wiggins 2001, 8–11, and n. 2 on 79.

> A concept of x is a body of knowledge about x that is stored in long term memory and that is used by default in the processes underlying most, if not all higher cognitive competencies when these processes result in judgments about x.
>
> (Machery 2009, 12)

Both characterizations are meant to be in some sense fundamental. They should provide one with a preliminary understanding of what a concept is, compatible with many different theoretical views. But they are strikingly different. We may individuate two main differences. Firstly, in Machery's characterization concepts consist of "knowledge". Even if in Machery the term 'knowledge' is used in a deflationary way, as roughly synonymous with 'information', or even 'true belief', Burge's concepts do not consist of knowledge. Knowledge requires a propositional content, but concepts for Burge are only subcomponents of thoughts; they do not have propositional content. Secondly, Burge's concepts aim to apply or refer. There is no mention of such categories in Machery's characterization. The concept (the body of knowledge) is used in reaching a judgment, but it is not even clear if it would make sense to say of a body of knowledge that it literally 'refers' to an object.

For the reasons explained above, I take Burge's characterization of concepts to be correct,[9] and I take Machery to be wrong. But Machery's characterization would be roughly correct if he were talking about conceptions. Does this provide a sufficient reason to say that psychological theories are really about conceptions, or ways in which we apply concepts? There is a sense in which what psychologists are really interested in is indeed the ability to apply concepts. However, this does not warrant by itself the suggested semantic reinterpretation. There is an alternative way of describing the situation, which is at least equally natural, and that is to say that often psychologists mistakenly identify the ability to apply a concept with the concept itself. Their theories provide information about conceptions, but they contain the mistaken assumption that conceptions are concepts. Here is an analogy: Columbus could report to his fellow Europeans a few true things about America, but in some sense he did not know what place he was talking about. Suppose he called the place he had reached 'China'. This description of his situation does not warrant interpreting his word 'China' as meaning what we mean by 'America'. In fact, given his beliefs about the size of the earth, it is much more natural to say that he was convinced he had reached China. Again, nothing I say here hangs on assimilating psychologists to the confused Columbus. But nothing excludes the possibility that they are confused, and that they provide us with some rudimentary maps of the ability to apply a concept, even though they take themselves to be investigating concepts *simpliciter*.

9 Of course that's not a *theory* of concepts. Most philosophically prominent theories of concepts are going to be compatible with that initial characterization. For example, causal-informational theories, Dretske 1981, and Fodor 1987 and 1998; teleological theories, Millikan 1984 and 2000); interpretivist theories, Davidson 1984 and Williamson 2007; and, to cite a recent proposal, the *originalist* theory of Sainsbury and Tye 2012.

My aim was to show what form the ability to apply a concept could take, respecting two requirements: the account should be compatible with what we currently know about the mind, and it should allow the ability to apply concepts to deliver reliable judgments about thought experiments. Clearly, the theories that I have described respect the first requirement. How about the second? It seems that they fully respect it. I am not trying here to show that philosophers are usually reliable in their judgments on thought experiments, and a fortiori I am not trying to explain how we are reliable, for example by giving a specific account of our conception of certain crucial notions. These are clearly questions that require separate consideration, and several empirical factors will be involved. What I am saying is that the proposal offers a model of how judgments on thought experiments *could* be reliable, without appealing to notions such as intuition or analyticity. This will be the case when our conceptions are sensitive to essential features of the phenomenon under consideration. Let me illustrate, with the paradigmatic case of thought experiment provided by Gettier cases. Our prototype or implicit theory of knowledge clearly could predict some kind of reliable connection between the subject's belief and the truth of the proposition involved. This requirement is typically satisfied when we obtain knowledge by perception, and in many other common cases that we classify as knowledge. The lack of this reliable connection in the belief involved in the Gettier cases would then yield a negative verdict; the lucky belief is not knowledge. On both the theories considered, there is no need for the subjects to be explicitly aware of what is involved in their capacity to apply a given concept. Thought experiments therefore can serve the purpose of making such information explicit. This is also clearly reminiscent of what Sorensen (1992) calls the 'recollection model' of armchair inquiry, a model going back at least to Plato. On this model, in considering new cases, actual or hypothetical, we tease out our own view of the subject matter involved, a view we already possess but need to make explicit.

The theories I described fully meet the worry that we were appealing to a mystery to explain a mystery. On the contrary, we are appealing to relatively well-understood mechanisms to explain a mystery. Recall my thesis SCV: A judgment about a thought experiment constitutes knowledge if and only if it is correct because it manifests a competence in applying the relevant concepts. Boghossian objected that we have no idea what a competence in applying a concept is. My reply is that a competence in applying a concept can be constituted by a reliable conception. So we can make SCV more informative by enhancing it as follows, with the addition of (ii):

Competent Conception View (CCV): (i) A judgment about a thought experiment constitutes knowledge if and only if it is correct because it manifests a competence in applying the relevant concepts—(ii) a competence in applying a concept is constituted by a reliable conception associated with the concept.[10]

10 Here, I take constitution to be distinct from identity; for reasons that will emerge in the next section, a competence cannot be identical to the conception by which it is constituted, although it could maybe be identical with the conception plus the history through which it was formed.

We are not appealing to *virtus cognitiva*, we are appealing to independently plausible mechanisms which have been investigated empirically for a long time, and we have not of course appealed anywhere to analytic connections or intuitions in the sense of a special mental state. In the next section, I will say something more about the consequences of the proposal, in particular with respect to the epistemic properties of the judgments.

4 Learning to apply concepts

In this section, I will further explain several aspects of the view I am proposing. I will start by considering an objection, which will lead me to clarify some features of my proposal. Then I will end by comparing my view to a similar one advanced by David Papineau.

Tim Williamson objected[11] to a previous version of this proposal, one restricted to prototypes, that although it seems plausible that the prototype plays a role in the ability to apply a concept, the latter cannot be identified with the prototype, for two people could have the same prototype but different abilities to judge how similar something is to the prototype. I think the point can be fully accommodated by two kinds of consideration. First, we have to take into account the competence–performance distinction. Certainly two subjects might have the same ability to apply a concept, while one of them reaches more correct judgments.[12] For instance one of the subjects might be often drunk, or generally unreflective and overconfident, and so on. Once all factors affecting performance are ruled out, however, the subjects will necessarily reach the same judgments. The prototype is not (just) an image that we have in the mind, which we have to compare to different objects. To associate a certain prototype with a concept requires being disposed to judge accordingly, in ideal conditions. On the view I am suggesting, the competence in applying the concept simply includes competence in judging how similar something is to the prototype. Similarly for having a tacit theory. Having a tacit theory will entail being disposed to judge according to it, once performance errors are ruled out. For any way of specifying any ability, one can always in principle ask about our ability to make use of the ability (the competence in moving from competence to performance); I see no particular worry about the specific form of CCV.

However, there is a second respect in which two subjects who share a prototype can differ, and that is the strictly epistemic respect. While two subjects who have the same prototype and perform equally well will reach the same judgments, the epistemic status of those judgments might differ. The point is more easily illustrated by considering the ability to apply a concept understood as an implicit

11 Pers. commun.

12 A further reason why two subjects with the same conception might reach a different number of correct beliefs is that those beliefs are formed in ways that involve directly some other belief-forming method, such as perception, testimony or episodic memory. In those cases, the reliability of the conception has to be judged conditionally on the correctness of the input they take.

theory. Our theories are not typically innate (even when we have an innate implicit theory, it typically undergoes some development). Old beliefs can be abandoned and new ones can be added, responding to experience and general revision processes. Now, this means that there is a sense in which an implicit theory might well be epistemically defective, even when it produces a correct belief. The beliefs that constitute the ability might fail to be justified, in the broad sense of having been formed as the result of an appropriate response to the empirical evidence or other rational considerations. So the following is a genuine possibility: two subjects have the same implicit theory, and perform equally well in applying it, thereby producing the same ratio of true beliefs, but the resulting beliefs constitute knowledge for one and not for the other. This would be possible when, in the bad case, the judgments are driven by beliefs that are not themselves in good epistemic standing, so that the conception, even if reliable, fails to constitute a competence.

A similar conclusion can be reached if conceptions are thought of as prototypes. It is uncontroversial that prototypes associated with a concept can change across cultures, across individuals, and across different times for even a single individual. Of course if we were to identify the prototype with the concept, this would entail that there are different concepts; but since we are now thinking of a prototype as an ability to apply a concept, the variation in prototypes only entails a variation in such an ability. Furthermore, although it is not clear whether we can say that the prototype itself is justified or known (for it does not consist of beliefs), we can surely say that one prototype is more reliable than another. More importantly for my present point, we can say that some prototypes are formed in epistemically better ways than others, e.g. responding appropriately to interaction with the world.

To sum up, in order for the conception to constitute a competence in a normative sense, it is not sufficient that it actually provides a sufficiently high ratio of correct judgments. The way the conception was acquired is also essential to its epistemic effects. This means that a strong form of epistemic externalism holds here; two individuals internally identical, who possess the same concepts and inhabit at present the same environment, may differ in the epistemic properties that their reasoning instantiates, because of a difference in their causal history.[13]

There is a further consequence of this picture of the conceptions underlying a competence in applying a concept, and of the way they are acquired and modified, which is worth highlighting, and is potentially of great importance for the self-understanding of philosophy. On the view I am defending, a judgment we reach about a thought experiment by making use of our competence in applying a concept, or even making use exclusively of that competence, need not be a priori or analytic. Typically, the competence required to reach the judgment will exceed what is required for concept possession, and it will depend on past

13 I do not rule out the existence of interesting epistemic properties that supervene on the totality of the subject's present mental states, but I do not see them as relevant in this context.

experiences for its epistemic standing.[14] In other terms, my reply to the *virtus cognitiva* objection provides reason, other things being equal, to favour SCV or CCV over RCV.

Papineau (2009) also argues for the view that judgments about philosophical thought experiments are the product of a capacity to apply concepts, and that such capacity encapsulates (to use his term) empirical information. I am of course also defending this general claim. However, there are some rather important differences between Papineau's account and mine[15] (besides the fact that my account goes a little further than his in describing the capacity itself). I will end by commenting on these differences.

Papineau thinks that "substantial synthetic assumptions are built into the automatic mechanisms that allow us to make particular judgments about philosophically salient categories like knowledge, names, persons, free will and so on" (Papineau 2009, 18). He also argues that this supports the view that the relevant judgments are a posteriori justified. I am sceptical about the usefulness of the category of a posteriori justification, just as much as I am sceptical about the usefulness of the category of a priori justification, but in this context this can be put aside. I am in agreement with Papineau that experience plays some role in shaping our capacity to apply concepts. But I am in disagreement with him where he talks about an "automatic mechanism". Papineau thinks the relevant capacity always operates at a subpersonal level. He compares the working of the relevant capacity to the way in which the visual system computes the shape of a represented object starting from sharp changes in intensity in the stimulation of the retina. I have talked about 'implicit' or 'tacit' theories. But this is not the same as situating the theories at a subpersonal level. A belief that most readers will share is that they are not giant pink giraffes. This is a tacit belief, but it is held at the personal level, and it is easily retrievable (different tacit beliefs might be less easily retrievable). By contrast, beliefs about the relation between the stimulation of the retina and the objects we are seeing are not retrievable, and they are not attributable to the subject. I do not want to rule out the possibility that some automatic, subpersonal mechanism is involved in the capacity to apply concepts. But I do want to deny that *only* automatic, subpersonal mechanisms are involved. On my view, the mechanisms involved may range from completely subpersonal to completely conscious, with all the intermediate degrees.

Papineau believes that describing the ability to apply concepts as an automatic mechanism helps in explaining the appearance that judgments about philosophical thought experiments are not falsifiable. I do not believe there is any such appearance, at least not a strong one. Judgments about thought experiments are often uncertain. The "sub-personal picture" of the capacity to apply concepts leads Papineau to a moderately pessimistic view of the epistemology of judgments about cases. Later on in the paper he writes:

14 A similar conclusion is reached, through a different route, in Williamson 2007, 165–169.

15 I should note that Papineau claimed, at the conference this volume derives from, that he does not currently hold the part of the view expressed in his 2009 that I criticize.

> The function of cognitive mechanisms that embody encapsulated assumptions is to deliver judgments about particular cases quickly and efficiently. Because of this, the relevant assumptions are standardly rules of thumb that work well enough in most cases but are not strictly accurate, in the way illustrated by the familiar perceptual examples. If the cognitive mechanisms behind philosophical intuitions are at all similar, we should expect encapsulated philosophical assumptions to have a similar status. They may work well enough for practical purposes, but they may not be strictly accurate and may lead us astray in certain cases. If we are to be confident about these assumptions, we will need to make them explicit and subject them to proper a posteriori evaluation.
>
> (Papineau 2009, 21)

As I said above, I do not find Papineau's motivations for his description of the capacity underlying hypothetical judgments very convincing. The resulting form of moderate scepticism is also worrying. Moreover, Papineaus's picture is in tension with recent developments in psychological research. A kind of view which enjoys growing popularity in the psychology of reasoning is represented by "dual-process theories". According to this kind of view, humans possess two reasoning systems, often labeled system 1 and system 2 (see, e.g., Frankish and Evans 2009). System 1 is, very roughly, evolutionarily old, unconscious, automatic, fast, based on associations. System 2 has the opposite features: evolutionarily recent, conscious, voluntarily controlled, slow, based on logical reasoning. At first glance it might seem that this conforms very well with Papineau's view. He is claiming that "intuitive" judgments proceed from system-1 reasoning. However, proponents of dual-process theories clearly hold that hypothetical reasoning triggers system-2 reasoning, the kind that is exactly opposite (Evans 2007 offers a detailed account of this claim, as well as much empirical support).

I believe that, if one considers the matter carefully, the view that philosophers typically form judgments about thought experiments through mechanisms *completely* inaccessible to conscious reflection ought to strike one as implausible on commonsensical grounds already. Philosophers spend large amounts of time in careful reflection about hypothetical cases. We should not assume that this is a waste of time without good reason. In addition, philosophers who have been practising the discipline for some time should have first-hand experience of the malleability of judgments about thought experiments. I find that my judgments at least are susceptible to change over time, sometimes due to theoretical considerations. To give just one example, Grice and Strawson (1956) ask us to imagine someone asserting 'my neighbour's three-year-old child is an adult'. They claim that the content of the assertion is a logical impossibility, and that if the asserter insisted that the claim has to be taken seriously and literally, we would not understand what they were saying. At first, despite my sympathy for the Quinean stance, I thought these judgments very plausible. At some point, however, it occurred to me in this connection that a three-year-old dog is an adult dog, and a three-year-old chimpanzee is an adult chimpanzee, and so on for all

other species I could think of, except one. Surely it is a rather interesting biological fact, and in a certain sense even a surprising and peculiar fact, that a three-year-old human is *not* an adult human. But then it seems the assertion above does not express a logical or epistemological impossibility, and not even a metaphysical impossibility (unless laws of nature are metaphysically necessary).[16] Whether or not I was correct in my change of mind on this case, the crucial philosophical point is that in considering a hypothetical scenario we allow ourselves to form judgments on the basis of all of our background knowledge, implicit and explicit, and whatever its origin, unless it contrasts with the assumption that the scenario holds.

5 Conclusion

I have argued that we can, in principle, gain knowledge through the use of thought experiments by employing no special faculty and no capacity essentially different from the ones required to yield judgments about actual cases. In particular, I have argued that we can gain such knowledge through a competence in applying concepts, and that some psychological theories of concepts, the prototype theory and the theory-theory, can be reinterpreted as providing a model of such a competence.[17]

References

Armstrong, S., Gleitman, L. and Gleitman, H. 1983. What some concepts might not be. *Cognition*, 13 (3): 263–308.

Barsalou, L.1987. The instability of graded structure: Implications for the nature of concepts. In U. Neisser (ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.

Boghossian, P. 2009. Virtuous intuitions: comments on lecture 3 of Ernest Sosa's *A Virtue Epistemology*. *Philosophical Studies*, 144: 111–119.

Brown, J. R. and Fehige, Y. 2011. Thought experiments. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2011 edition), <http://plato.stanford.edu/archives/fall 2011/entries/thought-experiment/>.

Burge, T. 1993. Concepts, definitions and meaning. *Metaphilosophy*: 24: 309–325.

Carey, S. 1991. Knowledge acquisition: enrichment or conceptual change? In S. Carey and R. Gelman (eds.), *The Epigenesis of Mind: Essays on Biology and Cognition* (pp. 257–291).

16 There is a sense of 'child' in which being a child and an adult is indeed impossible. There is also a sense of child in which it is not ("when children leave home, parents can feel somewhat redundant", to use *OED*'s example). The surrounding text of Grice and Strawson's claim suggests that it is not (just) a child being an adult that they consider logically impossible, but rather a three-year-old being an adult.

17 I wish to warmly thank the audience and the organizers of the Philosophical Insights conference for comments on the presentation of this material. In addition, I wish to thank Jessica Brown, Herman Cappelen, Lars Dänzer, Jens Kipper, Daria Mingardo, Ernest Sosa and Tim Williamson, for their help in making the paper better in form and content.

Hillsdale, NJ: Lawrence Erlbaum Assoc. Repr. in Margolis and Laurence 1999a: 459–487.

——2009. *The Origin of Concepts*. Oxford: Oxford University Press

Cummins, R. 1998. Reflections on reflective equilibrium. In DePaul and Ramsey 1998: 113–127.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press

DePaul, M. and Ramsey, W. (eds.) 1998. *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press

Evans, J. 2007. *Hypothetical Reasoning: Dual Processes in Reasoning and Judgment*. Abingdon: Psychology Press.

Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

——1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

——2008. *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Frankish, K. and Evans, J. 2009. The duality of mind: an historical perspective. In J. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 1–29). Oxford: Oxford University Press.

Gelman, S. and Wellman, H. 1991. Insides and essences: early understanding of the non-obvious. *Cognition*, 38 (3): 213–234. Repr. in Margolis and Laurence 1999a: 613–637.

Grice, P. and Strawson, P. 1956. In defense of a dogma. *Philosophical Review*, 65: 141–158.

Laurence, S. and Margolis, E. (eds.) 1999a. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

——1999b. Concepts and cognitive science. In Margolis and Laurence 1999a: 3–81.

Machery, E. 2009. *Doing without Concepts*. Oxford: Oxford University Press.

Malmgren, A.-S. 2011. Rationalism and the content of intuitive judgments. *Mind*, 120: 263–327.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

——2000. *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.

Papineau, D. 2009. The poverty of analysis. *Aristotelian Society Supplementary Volume*, 83: 1–30.

Rey, G. 1983. Concepts and stereotypes. *Cognition*, 15 (1–3): 237–262. Repr. in Margolis and Laurence 1999a: 279–299.

——2009. Review of *Doing without Concepts*, by Edouard Machery. *Notre Dame Philosophical Reviews*, 15 July, <https://ndpr.nd.edu/news/24087-doing-without-concepts/>.

Rosch, E. 1973. Natural categories. *Cognitive Psychology*, 4: 328–350.

——1978. Principles of categorization. In E. Rosch and B. B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum. Repr. in Margolis and Laurence 1999a: 189–206.

——1999. Reclaiming concepts. In R. Núñez and W. J. Freeman (eds.), *Reclaiming Cognition: The Primacy of Action, Intention and Emotion*. Thorverton: Imprint Academic.

Rosch, E. and Mervis, C. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 8: 382–439. Repr. in DePaul and Ramsey 1998.

Sainsbury, M. and Tye, M. 2012. *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts*. Oxford: Oxford University Press.

Sorensen, R. A. 1992. *Thought Experiments*. Oxford: Oxford University Press.

Sosa, E. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*, vol. 1. Oxford: Oxford University Press.

——2009. Replies to commentators. *Philosophical Studies*, 144: 137–147.

Wiggins, D. 2001. *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.