Philosophy Unbound

Michael Strevens

To appear in a *Philosophical and Phenomenological Research*
symposium on Edouard Machery's *Philosophy Within Its Proper Bounds*

1. The Method of Cases

Philosophy within its Proper Bounds takes aim at that paradigm of armchair philosophical reasoning, the "method of cases", in which hypotheses about some category or property's nature are tested against judgments about category membership or property instantiation. What makes a certain substance a specimen of water (in the everyday sense of that word)? It is clearly important that it contain quite a bit of H_2O. It is allowed to contain impurities, however—tap water is not 100 % H_2O, and seawater is typically 97 % H_2O or less. Is there a cutoff point above which everything with that proportion of H_2O counts as water? Evidently not: a reasonably strong cup of coffee is nearly 99 % H_2O, but it is not a cup of water. Perhaps flavor and color, then, can disqualify an otherwise suitable candidate for waterhood? Yet I have drunk much coffee that did not taste of coffee at all. It tastes like water, I thought. But I didn't think it was water.

Armchair philosophers hope to use the same strategy to analyze philosophically more interesting categories than water: to discover, by way of case judgments or "intuitions", the nature of knowledge, of singular causation, of justice. (I'll talk about categories only from now on; it should be easy enough to see how properties, such as justice, fit into the picture.)

2. The Argument from Unreliability

In this critique, I will take on just one of the arguments Machery offers against the value of the method of cases: the *argument from unreliability*, which turns on the claim that case judgments are in certain crucial scenarios for systematic reasons unreliable. Machery infers unreliability from features of case judgments that can be directly assessed through experimental inquiry, namely their heterogeneity and their volatility.

A class of case judgments exhibits heterogeneity if different groups of people judge differently about the cases. Machery's leading example, based on extensive cross-cultural testing, is the class of scenarios exemplified by Kripke's well-known Gödel case. Gödel is widely believed to have proved the incompleteness theorem that bears his name, but it was in fact (so the story goes) proved by Schmidt. Does the name "Gödel" refer to the person who really proved the theorem—the person who satisfies the description associated with the name—or rather the person who was given that name at birth? It seems that East Asians are inclined to give the former answer, Americans the latter. Even putting aside culture, there is considerable variation within each group: about 40 % of Asians give the "Kripkean" judgment about scenarios of this sort, whereas about 40 % of Americans give the "descriptivist" answer. As Machery points out, it seems rash, if not positively reckless, to depend on a judgment with respect to which 40 % of the population disagrees with you (given that there is no ideological or other reason to suppose that the minority are inferior adjudicators of these matters). Yet this very judgment has played a crucial role in the philosophical analysis of reference.

Other research suggested at first that there were cultural variations in judgments about Gettier cases. That finding has not held up, yet work in the same vein shows that there is considerable variation in Gettier judgments within apparently homogeneous groups. Turri et al. (2015), for example, find their respondents split almost down the middle when it comes to the

attribution of knowledge in what appears to be a fairly classic Gettier scenario.[1]

A class of case judgments exhibits volatility if the same person judges differently about the cases in the class in different circumstances or when evidently irrelevant aspects of the case are varied. Valdesolo and DeSteno (2006) found, for example, that subjects presented with the standard trolley problem were more likely to push the "fat man" onto the tracks if they had first viewed a short comedy sketch. And Swain et al. (2008) showed that subjects are more likely to attribute knowledge to "Mr. Truetemp"—a person who has infallible beliefs about the ambient temperature without understanding why—if they had previously been asked to make a judgment about a case in which a person clearly lacks knowledge as opposed to a case in which they clearly possess it.

In *Philosophy within its Proper Bounds*, Machery marshals numerous examples of heterogeneity and volatility to argue that in many of the scenarios crucial to the success of the method of cases, case judgments are unreliable. Most of our case judgments, he acknowledges, are correct; yet in just those special scenarios that promise to render a verdict between competing analyses of philosophical matters such as reference and knowledge—in Gödel cases, Gettier cases, and so on—our minds go mushy and fail to deliver a reliable verdict. It is as if in scientific inquiry, it turned out that precisely when we staged a crucial experiment, the needles on our instruments began to waver and tremble. That would be enough to stymie the enterprise of science, and it is equally enough, Machery contends, to undercut any warrant we have for relying on the method of cases.

In order for Machery's argument to succeed in undermining our trust in the method of cases, our unreliability must be systematic and incurable. Were

1. The case is as follows: Emma purchases a diamond from a jewelry store and puts it in her pocket. A skilled jewel thief tries to steal it from her pocket before she leaves the store, and he succeeds. Someone secretly slips a diamond into Emma's pocket before she leaves the store. About 55 % of Turri et al.'s respondents agreed that Emma knows she has a diamond in her pocket.

subjects on further consideration to retract their attributions of knowledge in Gettier cases, for example, resulting in a near unanimity among careful deliberators that Gettierized justified true belief is not knowledge, then the contrary views of hastier or less attentive thinkers would be of little concern. Likewise, if Truetemp judgments were consistent outside of deliberately manipulative contexts, we could deal with the concerns raised by Swain et al. like any scientific experimenter, carefully conducting our thought experiments in isolation from distracting or noisy backgrounds.

The research surveyed in Machery's book provides, at the very least, some reason to think that heterogeneity and volatility are systematic. The question I want to ask in the remainder of this piece is whether that implies incurable unreliability.

The backdrop is the human mind, and more particularly the psychology of concepts, such as the concept of water, the concept of knowledge, or the concept of justice. Before taking on the unreliability argument, then, let me review two distinct views of the psychology of concepts, the classical and the theory-theory views.

3. The Psychology of Concepts

According to the classical theory, attached to any concept is a mentally represented definition that supervises the deployment of the concept in thought. The concept of knowledge, for example, is associated with a definition of knowledge that rules over our every inference about and judgment of knowledge (though perhaps at a distance, functioning more like a court of appeal while handing off everyday thought to various heuristics).[2]

Because it is a definition, this criterion reflects, indeed stipulates, the nature of the corresponding category. It follows that our categories' natures are represented in the head; to determine those natures is therefore to perform

2. On both the classical theory and theory-theory, see Margolis and Laurence (1999).

"conceptual analysis". But it is not as easy as it sounds, because we cannot determine the structure of these definitions, or most of them, by direct introspection. We must use something like the method of cases to figure out their contents.

The classical theory explains a great deal about the character and power of the method of cases. When we think carefully and deliberately, we are applying our mental definitions to the philosophical thought experiments that elicit case judgments. Consequently, case judgments are privileged. Confronting a well-designed thought experiment and thinking clearly, we cannot go wrong. There lies the security and force of philosophers "intuitions" about cases.

Further, such judgments, used imaginatively, will in time reveal the structure of the definitions that guide them and so will reveal the natures of the categories concerning which the judgments are made. My mental definition of water will eventually disclose itself if I make sufficiently many waterhood judgments about sufficiently diverse substances. The same goes for my mental definition of knowledge. It is a shame that I cannot simply peruse the pages of my mental dictionary to read off the definitions, and so the natures, that are inscribed there, but the method of cases is, if far slower and more effortful, in the long run equally reliable.

Unfortunately for this age-old and reassuring conception of the workings of armchair thought, very few psychologists and philosophers today believe that the classical theory correctly describes more than a handful of concepts (such as "prime number"). Its sleekest rival, which is the view to which both Machery and I subscribe, is the so-called theory-theory of concepts.

The theory-theory holds that concepts, or most of them, correspond to something like theories, if only incipient ones, of the things in question—of the categories in question, I will say, since I am focusing on concepts of categories. Our concept of water is a (perhaps rudimentary) theory of water, consisting of hypotheses such as "water is transparent" and "water is conductive". Or as Machery glosses it, concepts are "bodies of information", "belief-like states",

"about individuals, classes, substances, or events" (*Philosophy within its Proper Bounds*, 210).

Two important remarks about these theories. First, the elements of such a theory are entertained in a merely suppositional spirit, and may be abandoned if evidence mounts against them. My theory of cats may contain the hypothesis "Cats are animals"; were I to uncover the right sorts of facts I might relinquish this belief and form another to the effect that "Cats are robots".

Second, a theory need not contain a thesis about the corresponding category's nature. Arguably, even sophisticated scientists entertain theories of this sort: a physicist who knows a lot about muons may have no view about the metaphysics of muons, and *a fortiori* no view about the nature of muons. Thanks to their theoretical knowledge, they are a muon expert, and they are especially expert in identifying any muons that happen to fly by, but they do so using ordinary empirical knowledge about muons—their characteristic effects, their means of generation, and so on—that in no way purports to spell out what it is to be a muon. In my view, ordinary thinkers' concepts—and I include scientists among my ordinary thinkers—almost never include hypotheses about, let alone definitions of, natures. Even an expert botanist, for example, is not a plant metaphysician, but rather owes their expertise simply to knowing a large number of important things about certain classes of plants—about their characteristic appearances, patterns of growth, internal structure, and so on. That is enough for highly accurate categorization, among many other important cognitive tasks.

4. Is Incurable Unreliability Possible?

On the classical view of concepts, it is hard to see how we could be, as Machery claims, incurably unreliable. Our case judgments are supervised by mental definitions, and those definitions by their very nature cannot go wrong. We could perhaps be transiently unreliable if the application of a definition were complicated or tricky in a certain range of cases, but such unreliability could

surely be rectified with close attention and some inferential chops.

On the theory-theory of concepts, by contrast, irremediable unreliability seems to be a real possibility. Judgments of category membership are made, the theory-theorist holds, using hypotheses that collectively amount to "just a theory". Individual hypotheses do not have the privilege accorded to them by the classical theory, which supposes that they are made using an apodictic criterion for category membership, a definition. Thus they could be false. Or even more plausibly, they could be incomplete, supplying less than the reasoner needs to find their way to a secure judgment about the scenario in question.

In this way, the theory-theory opens up a gap between a category and its case judgments that would not exist if the classical theory gave the correct account of philosophical concepts such as knowledge, causality, and so on—a gap that allows for the possibility that philosophical case judgments are in some instances unreliable.

Why should the paradigmatic philosophical thought experiments prove particularly problematic? Because, Machery suggests, in order to distinguish between otherwise plausible theories of natures these scenarios "pull apart what usually goes together" (§3.5.4). In a typical case of knowledge, the reason a believer is correct is also the reason that they are justified; a Gettier case deliberately separates the two. In a typical case of historical reference, what we believe about a person baptized with a certain name is largely true of the person, or at least not true of one of their contemporaries baptized with a different name; Kripke's Gödel case pulls beliefs and baptism apart.

Machery offers four possible stories why pulling apart should stretch our cognitive apparatus to breaking point, resulting in unreliable case judgments. I consider two of them here. The first draws on an idea discussed in earlier work (Machery and Seppälä 2011). According to this picture, any given concept is typically connected to multiple competing criteria for category membership, with no master criterion to arbitrate when they disagree. In straightforward

cases, the criteria agree, and so case judgments proceed without difficulty. In the classic thought experiments, by contrast, different criteria suggest different judgments: a Gettier case might qualify as knowledge according to one criterion but not according to another. Because there is no principled way of adjudicating such a clash, judgments become subject to various outside forces: cultural norms, professional ambition, immediate context.

If the theory-theory of concepts is correct, this picture is not quite right. True, an ordinary thinker's mental theory may very well supply various criteria for category membership that occasionally disagree, just as an incomplete scientific theory may supply various tests for the presence of invisible objects or properties that occasionally disagree. But because these criteria are not isolated, but are rather parts of a single body of information, their verdicts may be weighed against one another by using our standard rules of inference, and above all, by inductive logic. If I look out the window and think I see rain, but my phone tells me that the sun is shining, I am not helpless. I can think about which source of information is more reliable in the context, and favor one verdict decisively over the other, reaching a stable judgment about the local weather.

That said, as all inductive reasoners know, it is far from unusual for inductive considerations to yield nothing more helpful than a shrug of the shoulders. Such equivocation lies at the heart of another of Machery's explanations for unreliability, according to which there are multiple criteria for category membership that sometimes deliver conflicting judgments, and the master criterion that we have for adjudicating among such judgments is unable in many cases to make a clear ruling. If the theory-theory of concepts is correct, such predicaments should not be at all rare.

The average person confronted with a Gödel case, you might speculate, experiences a train of thought of roughly the following character. One part of them wants to say that "Gödel" refers to the person baptized with that name. One part of them wants to say that it refers to Schmidt. Neither part seems

clearly more authoritative than the other, so the usual inductive techniques for reaching a conclusion are unhelpful. Enter culture, circumstances, context . . . and thus, in cases where a clear verdict is demanded, heterogeneity and volatility.

Although I will suppose that this explanation is largely correct, I will attempt to undermine Machery's unreliability argument against the method of cases by resisting the inference from heterogeneity and volatility to unreliability. I have two lines of attack.

The first is to observe that theories can normally be expected to improve over time. When I was five years old, my concept of pine trees was attached to a very crude theory—a set of beliefs about pines as scrawny as my five-year-old physique—and so I was a rather unreliable classifier of that kind of tree. As I grew older and learned more, my theory filled out and I can now usually tell a pine when I see one. Machery's argument requires unreliability to be incurable, but thinness or falsehood in a theory can be rectified. We shouldn't let just anyone sit in the armchair, perhaps, but it is safe to allow entry to philosophical experts.

And yet: is there really such a thing as a philosophical expert? It is easy to improve your knowledge about pines, but it is not so obvious how to improve your knowledge about philosophical matters such as singular causality or, indeed, knowledge. Certainly, scientific findings do not seem to have much bearing on judgments about Gettier cases. Perhaps we can improve our philosophical concepts through reflection? I don't mean the method of cases, whose efficacy is precisely the issue at hand, but processes such as the removal of inconsistencies. In this way, a philosopher's concept might be more unified than an ordinary person's, having few or no criteria for category membership that "pull apart". It is far from clear to me that we can improve our theories in this way, but if the goal is to incapacitate Machery's argument from unreliability, as opposed to composing an independent defense of the method of cases, perhaps it is enough to point to the possibility. Unreliability

now, especially among novice thinkers, does not mean unreliability forever.[3]

My second strategy for subverting the argument from unreliability is to ask whether there can be a fact of the matter about category membership in cases where we have a systematic, incurable inability to make a membership judgment. I suggest that membership in such cases is indeterminate, and thus that our failure is not an instance of unreliability but rather of something that poses far less danger to the method of cases.

Suppose that concerning a certain scenario, human judgment is volatile for the reason suggested above: ordinary human theories of the category in question lack the resources to make a decisive call on category membership. Suppose further that this is one of Machery's crucial scenarios: sitting on the philosophical table are two plausible theories of the category's nature, one of which classifies the scenario in one way, and the other in the opposite way. One theory, that is, says that the scenario is an instance of knowledge, or singular causation, or moral responsibility, while the other says that it is not knowledge, not causation, not responsibility. If we are unable to reliably form an opinion about such cases and others with the same strategic leverage, Machery argues, then we are unable to use the method of cases to discover the category's nature.

But that is true only if there is a fact of the matter about membership. Suppose that there is none: it is indeterminate whether the scenario in question exemplifies the category or not. Then both theories on the table must be mistaken, because both, by way of making definite classifications, deny indeterminacy. The true nature of the category leaves membership undecided; the theories do not. Thus they are not theories of the true nature.

In this situation, there is still a deficit in our judgments: we find ourselves uncertain about category membership, rather than certain that membership is indeterminate. In that state of mind, we will be unable to recognize the inadequacy of the two theories on the table, and thus we will remain unaware

3. For my own attempt to vindicate the method of cases, see Strevens (2019).

of the need for a theory according which there is no fact about membership.

Over time, however, it ought to become clear to us that our failure to resolve these cases is systematic and incurable. Now we might say to ourselves: what could determine facts about category membership except the sum total of our concepts, patterns of thought, and behavioral inclinations? If there is nothing in the human world to decide the question of category membership one way or another, then there can be no fact of the matter. By this line of thought—inspired, I think, by the pragmatist or verificationist notion that our words go no further than we are capable of carrying them—we arrive at the conclusion that membership is indeterminate, and so come to discard theories that make determinate judgments in the cases in question.

To put it more concretely, if we were really unable to make up our collective minds as to whether a particular Gettier-like case were an instance of knowledge—say, the case where a thief attempts to steal Emma's diamond but fails—then we should conclude that there is no fact of the matter. Any theory of knowledge that says otherwise, that is, any theory that determinately classifies such cases as knowledge or as non-knowledge, should be rejected in favor of a theory that for principled reasons marks the status of Emma's belief as indeterminate.

This is reassuring, I hardly need to add, only if my pragmatist precept is correct, that is, only if systematic, incurable uncertainty or volatility about category membership implies indeterminacy. I find the precept to be a compelling one. But in order to undermine Machery's argument from unreliability, it need not be proven or widely held. It need merely be a real philosophical possibility.

References

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press, Oxford.

Machery, E. and S. Seppälä. (2011). Against hybrid theories of concepts. *Anthropology and Philosophy* 1:99–127.

Margolis, E. and S. Laurence. (1999). *Concepts: Core Readings*, chap. 1, pp. 3–81. MIT Press, Cambridge, MA.

Strevens, M. (2019). *Thinking Off Your Feet: How Empirical Psychology Vindicates Armchair Philosophy*. Harvard University Press, Cambridge, MA.

Swain, S., J. Alexander, and J. M. Weinberg. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research* 76:138–155.

Turri, J., W. Buckwalter, and P. Blouw. (2015). Knowledge and luck. *Psychonomic Bulletin & Review* 22:378–390.

Valdesolo, P. and D. DeSteno. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science* 17:476–477.