BRILL | MENTIS

HISTORY OF PHILOSOPHY & LOGICAL ANALYSIS
27 (2024) 370–394

History of Philosophy & Logical Analysis
Philosophiegeschichte und logische Analyse
brill.com/hpla

Reflecting on One's Own Philosophical Practice

Timothy Williamson | ORCID: 0000-0002-4659-8672
Faculty of Philosophy, University of Oxford, Oxford, United Kingdom
timothy.williamson@philosophy.ox.ac.uk

Received 22 March 2024 | Accepted 5 April 2024 |
Published online 13 January 2025

Abstract

Metaphilosophy is defined as philosophical reflection on philosophy itself, and so is part of philosophy. If one lacks the category of philosophy, one can still do philosophy, but one may not be in a position to do metaphilosophy. Cases are discussed of tension between a philosopher's metaphilosophical theory and their philosophical practice, in epistemology (scepticism), metaphysics (ontological minimalism), and philosophy of language (verificationism and intensionalism). Such tensions often provoke charges of self–defeat. However, the self–defeat turns out to be just one more manifestation of more general inadequacies in the philosophical theory at issue. The epistemology of philosophy is just an application of general epistemology; the philosophy of philosophical language is just an application of general philosophy of language, and so on. These considerations are used to support *anti–exceptionalism* about philosophy: philosophical thought and talk are much less different from other thought and talk than many philosophers suppose. The second half of the paper describes the author's own experience of the complex dialectic between metaphilosophy and more general philosophy, in which metaphilosophical disputes about specific aspects of philosophical practice draw attention to more general cognitive phenomena, which require to be understood philosophically, but may also call for revisions of philosophical practice in metaphilosophically significant ways.

Keywords

metaphilosophy – theory and practice – scepticism – verificationism – intensionalism – anti-exceptionalism – counterpossibles – conditionals – suppositional heuristic

1 Introduction

A philosopher–turned–metaphilosopher can ask: how is my metaphilosophy related to my philosophy? This paper reflects on that question. Although it is not intended as a direct contribution to the history of metaphilosophy, it raises some general issues which may serve as running threads through that history. In effect, it is an essay in metametaphilosophy. It may provide an historian of metaphilosophy with some questions to ask about ways in which their target philosophers' metaphilosophical ideas related or did not relate to their philosophical but non–metaphilosophical ideas.

At a first pass, one may define 'metaphilosophy' as philosophical reflection on philosophy. Thus, metaphilosophy is part of philosophy, not something above or beyond it, as I have used the term 'the philosophy of philosophy' to emphasize (Williamson 2007). One's metaphilosophy is related to one's philosophy at least as part to whole.

One might qualify the first–pass definition by saying: metaphilosophy is philosophical reflection on philosophy *qua philosophy*. For a philosopher studying the general nature of human thought or speech might reflect on an episode of philosophical thought or speech with no special attention to its specifically philosophical character, having picked it as an example of thought or speech simply because it came first to mind. Would that really count as metaphilosophy, or philosophy of philosophy?

Given the extra gloss on 'metaphilosophy', one can engage in metaphilosophy only if one has the category of philosophy, to apply to what one is reflecting on. Nevertheless, someone can still engage in *philosophy* itself without having the category of philosophy. We, who have the category, may still recognize their questions as philosophical questions, and their ways of answering them as philosophical ways, even though they, who lack the category, classify what they are doing as 'science' or 'literature', or 'wisdom', or just 'thought', and reflect on it as such. Self–identifying as a philosopher is neither necessary nor sufficient for *being* a philosopher. After all, treating it as such a necessary condition would risk making philosophy a parochial enterprise, by excluding other traditions which taxonomize the space of possible inquiry differently from the way we do.

Even within the philosophical tradition that traces back to ancient Greece, it is unclear how far back *our* category of philosophy goes. Etymologically, the *word* 'philosophy' and its cognates go all the way back, but that does not mean that they have always expressed our category of philosophy. Until the eighteenth century, inquiries which *we* classify as 'physics' were classified as 'philosophy' – 'natural philosophy'. They were 'philosophy' in the sense the

word had then, but that does not make them 'philosophy' in the sense the word has now. Consequently, metaphilosophy in the strict sense may be a comparatively recent development.

In order to avoid such a restrictive conception of metaphilosophy, one might therefore delete the gloss '*qua* philosophy', and return to the simpler definition of metaphilosophy as philosophical reflection on philosophy, irrespective of whether it is conceived *as* philosophy. In this essay, I will stay neutral on that taxonomic issue, but focus on the schematic case of a philosopher reflecting on their own practice, which is indeed a philosophical practice, whether or not they conceive it exactly as such.

2 Incoherence Between Theory and Practice in Philosophy

Reflection on one's own philosophical practice might be expected to serve as a means to achieve coherence between one's philosophical practice and one's philosophical theory: ideally, one's philosophical theory would *vindicate* one's philosophical practice. We may expect such coherence to be common in the history of philosophy.

We can be more specific about the predicted coherence by considering specific branches of philosophy, in particular: epistemology, the philosophy of language, and metaphysics. Even if the division of philosophy into such branches is comparatively recent, we may still project the distinctions retrospectively onto periods when they had not yet been made, provided that we are clear about what we are doing.

For epistemology, a natural prediction is that, in their philosophical practice, a philosopher will employ cognitive methods which their epistemology counts as *good methods* for achieving knowledge, or at least justified belief.

For the philosophy of language, a minimal prediction is that a philosopher's philosophical discourse will at least count as *meaningful* by the standards of their philosophy of language.

For metaphysics, an even more minimal prediction is that a philosopher will at least count as *existing* by the standards of their metaphysics.

No doubt these apparently undemanding conditions for coherence are often met in the history of philosophy. In this paper, however, we will consider some notorious ways in which the conditions have sometimes been violated. Such cases of severe strain between philosophical theory and philosophical practice may be more instructive about the constraints on philosophizing. Here are some examples.

For epistemology, sceptical traditions have struggled with the charge of self–defeat (Burnyeat 1983). At its simplest, the assertion 'Nothing is known' may be met with the question 'How do you know?' Of course, more sophisticated sceptics have learnt not to *assert* 'Nothing is known', indeed, not to assert anything at all. They have to insinuate doubt less directly. They may hope to draw dogmatists' attention to sceptical arguments to whose premises and validity the dogmatists themselves are already implicitly committed. Sceptics can put arguments on the table without endorsing them. Still, if they want to present scepticism as a live alternative, they must at least indicate an answer the question 'How do you live your scepticism without continually violating its own precepts?' If they say that they just live by the appearances, they must explain how they have cognitive access to the appearances. For even statements of the form 'It appears to me that P' are true or false. After all, they can be lies. Less obviously, even by non–sceptical standards, we are not always in a position to know whether things appear to us a given way, and if we guess, we may guess wrong (Williamson 2000). Sceptical attempts to live a life without risk of error look hopeless.

For the philosophy of language, a celebrated example is Wittgenstein's *Tractatus Logico–Philosophicus*, which is explicitly senseless by its own standard, since it consists largely of non–contingent statements, which fail to divide possibilities into those in which they are true and those in which they are false. Another familiar example is the logical empiricists' signature Verification Principle, that every cognitively meaningful statement is either tautologous or empirically verifiable (strictly speaking, such standard formulations of the principle apply only to *true* statements: they ignore the point that the negations of cognitively meaningful statements should also count as cognitively meaningful). Critics soon pointed out that the Verification Principle itself is neither tautologous nor empirically verifiable, so by its own standard it is cognitively meaningless. Classifying it as a mere stipulative definition of the term 'cognitively meaningful' did not help the logical empiricists, since by cutting the term off from its ordinary meaning the move rendered harmless their standard accusation that their opponents' discourse lacked 'cognitive meaning'.

For metaphysics, some philosophers argue that there are no macroscopic objects, or that no objects persist through change. Yet philosophers are people, and presumably people are macroscopic objects and persist through change. Consequently, such philosophers are in effect arguing against their own existence. Most explicitly, Peter Unger drew the anti–Cartesian conclusion 'I do not exist', on the grounds that, if he existed, he would be susceptible to a sorites paradox, and nothing susceptible to a sorites paradox can exist (Unger 1979). Similarly, if persons are selves, then Buddhists or Humeans who argue for a

no–self theory are in effect arguing that they do not exist. After all, any person can in principle reflect 'I am myself, so I am my self, so I am a self'. Philosophers who have proved their own non–existence to their own satisfaction cannot even fall back on the thought that, although they do not exist, at least their books still exist, for their books have the same features to which they object in themselves: books are macroscopic objects and persist through change.

The epistemological, semantical, and ontological humiliations that such philosophers heap on themselves can be effective strategies for winning converts, just like self–flagellation. They look like evidence of intellectual or moral integrity.

Philosophers' difficulties in making their philosophical theories and their philosophical practice cohere with each other might be taken to suggest the melodramatic conclusion that philosophy is an inherently self–defeating or paradoxical enterprise. But, to the contrary, on the supposition that philosophical progress *is* possible, one might ask oneself: why expect it to be easy? In particular, although philosophers are typically conscious of their own philosophical activity, that does not make its nature – especially its more general nature – transparent to them. It may be hard to construct a philosophical theory that does not mischaracterize that very process of construction.

One might be tempted to frame the problem by asking: why is metaphilosophy so difficult? Is philosophy a peculiarly elusive subject matter? Or is the challenge the reflexive, perhaps paradoxically self–referential nature of metaphilosophy?

Such suggestions are unpromising. Consider again the examples of self–undermining philosophical theorizing. Although the charge of self–defeat is dialectically elegant and effective, it does not capture what is most obviously implausible about the philosophical theories in question.

In epistemology, the most urgently problematic consequence of sceptical arguments is not just that they exclude knowledge *in philosophy* but that they exclude knowledge *in general*, especially non–philosophical knowledge, such as common–sense knowledge and scientific knowledge.

In metaphysics, the most urgently problematic consequence of no–self arguments is not that they imply the non–existence *of philosophers* but that the imply the non–existence of macroscopic and persisting objects *in general*, especially common–sense objects such as people and scientifically discovered objects.

In the philosophy of language, the Verification Principle is just one manifestation of a more general verificationist theory of meaning, on which the cognitive meaning of a declarative sentence is its canonically associated method of verification, and the most urgently problematic consequence of that general

theory of meaning is not the Verification Principle itself but the failure to identify methods of verification canonically associated with most declarative sentences of both ordinary and scientific language. For example, we can verify the sentence 'Jo is at home' in many different ways, but the English language does not privilege any one of them as somehow canonical. Similarly, we may be able to verify the sentence 'There are black holes' in many different ways, but again scientific English does not single out any one of them as canonical. The main obstacle is the holism of verification, emphasized by Quine (1951). In particular, the logical empiricists never provided any systematic way of determining the canonical verification condition for a complex sentence from the verification–related properties of its constituents, as the standard constraint of compositionality on a theory of meaning would require. Verificationism as an approach to the semantics of natural languages never got beyond the merely programmatic stage. Although applying verificationism to itself with respect to the Verification Principle is a dialectically effective move, what it highlights is just one specific manifestation of a far more general problem.

What about the other example above of self–defeat in the philosophy of language, the self–condemning account of meaning in the *Tractatus*? The case is subtler, but also revealing. It deserves a section to itself.

3 Case Study: Meaningful Impossibilities and the Individuation of Content

In order to avoid the contested bogs of Wittgenstein hermeneutics, I will focus on a contemporary approach to semantics. I suggest, but will not attempt to argue in detail, that the problem of self–defeat it faces is closely analogous to that for Tractarian semantics.

According to *intensional* semantics, in a given context a declarative sentence expresses a *proposition* which can be identified with a set of possible worlds, those at which the sentence is true; that set is the semantic value of that sentence in that context. Intensional frameworks are widespread in contemporary formal semantics, as practised in both the philosophy of language and linguistics. One notable defender of the intensional approach to both language and thought is Robert Stalnaker (1984; 1999).

A consequence of the intensional approach is that any non–contingent declarative sentence expresses either the set of *all* possible worlds, if it is true, or the set of *no* possible worlds, if it is false. Thus, given that sentences of pure mathematics are non–contingent, all true sentences in the language of

mathematics express the same proposition, as do all false sentences in that language. Likewise, if sentences of pure philosophy are non–contingent, all true sentences in the language of philosophy express the same proposition, as do all false sentences in that language. More specifically, any true sentence in the language of mathematics or philosophy expresses the proposition that 0 = 0 (or a trivial tautology), and any false sentence in that language expresses the proposition that 0 = 1 (or a blatant contradiction). The natural upshot seems to be that both mathematics and philosophy are cognitively insignificant. A contingent sentence divides possible worlds into some at which it is true and others, the rest, at which it is false, so we can have a serious inquiry into which side of the line the actual world is on. For a non–contingent sentence, there is no such distinction, and nothing to inquire into. Everything is either contingent and empirical, a Humean matter of fact, or non–contingent and merely conceptual, a Humean relation of ideas. Intensionalist philosophy seems to condemn itself to cognitive insignificance. By its own lights, it is either a tautology or a contradiction.

But appearances are deceptive. The insistence on the anomalous status of non–contingent sentences *underestimates* the generality of the problem. For the underlying issue concerns *non–trivial necessary equivalence*, and is just as serious for contingent sentences as for non–contingent ones.

Consider (1A) and (1B):

(1A) I own a ring made of gold.
(1B) I own a ring made of the element with atomic number 79.

On the metaphysical view of natural kinds powerfully defended by Kripke (1972; 1980), (1A) and (1B) are necessarily equivalent. Consequently, given intensionalism, the sentences (1A) and (1B) express the very same proposition. Yet their equivalence is far from trivial. Indeed, it seems obviously possible for someone ignorant of chemistry to know or believe (1A) without knowing or believing (1B). Yet (1A) and (1B) are highly contingent.

Similarly, consider (2A) and (2B):

(2A) There are 17^2 tiles on the floor.
(2B) There are 289 tiles on the floor.

By standard arithmetic, (2A) and (2B) are necessarily equivalent. Consequently, given intensionalism, the sentences (2A) and (2B) express the very same proposition. Yet their equivalence is far from trivial. Indeed, it seems obviously

possible for someone who has counted the tiles along the sides of the square floor but is bad at multiplication to know or believe (2A) without knowing or believing (2B). Yet (2A) and (2B) are highly contingent.

Thus, intensionalism undermines our natural judgments about cognitive significance just as much for contingent sentences as for non–contingent ones. To interpret the problem of self–defeat for intensionalist philosophies of language – including that of the *Tractatus* – as showing something about the status of non–empirical inquiries, such as mathematics and philosophy, rests on a radical misdiagnosis. Whatever it shows is something about cognitive significance for all language. The most urgently problematic consequences of intensionalism are not for philosophical discourse itself but for everyday and scientific discourse, as with (1A)/(1B) and (2A)/(2B).

It is tempting to treat the pairs (1A)/(1B) and (2A)/(2B) as straightforward counterexamples to intensionalism: if one can have an attitude to a proposition p without having it to a proposition q, then, by Leibniz's law of identity, $p \neq q$. That would have been Frege's reaction. For him, (1A) has the same reference (truth–value) as (1B), as uttered in a given context, but they differ in sense, and so express different thoughts. Thus, one can have an attitude of knowing or believing to the thought expressed by (1A) without having it to the thought expressed by (1B). He could take a similar line with (2A) and (2B). Such a Fregean line is also applicable in principle to mathematical and philosophical discourse: two sentences of mathematics or philosophy may have the same reference (truth–value) but differ in sense. That allows some sentences of mathematics and philosophy to express true cognitively significant thoughts.

On such a Frege–inspired view, the problem with intensionalism is its reckless attempt to do without a distinction between sense and reference. However, despite its early promise, the distinction has fallen out of favour in the philosophy of language since the 1970s. One reason is that Frege puzzles analogous to that of 'Hesperus' and 'Phosphorus' were shown to arise even for pairs of clearly synonymous terms, such as 'London' and 'Londres' or 'furze' and 'gorse' (Kripke 1979). Neo–Fregeans proposed that differences in sense can cut finer than differences in linguistic meaning for a given speaker, but if no two words can have the same sense, then the posited difference of sense becomes redundant; the difference of word will already do the work. The level of sense is like a useless layer of middle management between the level of syntax and the level of reference, ripe for cutting.

Another reason for the decline of Fregean semantics is that, even in the cases where it is most plausible to associate a word with a sense or mode of presentation of its reference, that mode of presentation does not make the

predicted contribution to the thought attributed to another thinker. For instance, consider (3):

(3) You think that I gave a talk at the Antwerp conference.

One might well associate the first–person singular pronoun ('I') with a supposed special way in which the speaker or thinker is presented only to themselves, as Frege did (in 'The Thought'). One might therefore expect this first–personal mode of presentation to be a constituent of the thought which I attribute to you when I utter (3). But that does not work, for when you employ the first–personal mode of presentation you refer to yourself, not to me, whereas the truth – condition of the thought which I attribute to you when I utter (3) is that *I* (TW) gave a talk at the Antwerp conference, not that *you* (dear reader) gave a talk at the Antwerp conference. The most straightforward application of the Fregean apparatus to (3) produces only a confusion quite irrelevant to an ordinary hearer's understanding of (3). With enough special pleading, Frege or his followers may be able to deflect (3) as a direct counterexample to Fregean semantics, but the point remains that, in a case with a salient candidate for the sense of the relevant singular term, that sense cannot be heard as a constituent of the attributed thought.

For such reasons, the sense–reference distinction does not really solve the problems it was intended for. No fine–graining of semantic content will do justice to all differences in cognitive significance. We must be willing to work with comparatively coarse–grained contents, such as intensions, in the compositional semantics, while allowing elsewhere for the cognitive significance of linguistic form – as in the lexical difference between two synonyms – and other factors, such as perceptual attention – as in the difference between two co–referring tokens of 'this rope', uttered while looking at opposite ends of the rope, perhaps tangled up with other ropes. For metaphilosophy, the upshot is that we cannot appreciate issues of cognitive significance in philosophy without tracking linguistic form as well as semantic content. The same goes for issues of cognitive significance in mathematics.

Does this mean that, in some subtle way, philosophy and mathematics are *about* language? That would be like the conclusion that in advancing from (1A) to (1B), or from (2A) to (2B), what one learns about is language. For example, in advancing from (1A) to (1B), I might go from knowing the contingent truth that the syntactic string (1A) expresses a true proposition in my language to knowing the distinct contingent truth that the syntactic string (1B) expresses a true proposition in my language; likewise for the advance from (2A) to (2B).

Stalnaker (1984; 1999) has in effect developed a treatment of cognitive significance along such lines within an intensional framework. However, it is not very plausible. In going from (1A) to (1B) or from (2A) to (2B), my cognitive focus is not on syntactic strings, and what they express; it is firmly on my ring, and what it is made of, or on the floor, and how many tiles are on it. Furthermore, the syntactic strings (1A), (1B), (2A), and (2B) are easy to parse and understand in my language; before making the advance, I already know what propositions they express in my language: (1A) the proposition that I own a ring made of gold, (1B) the proposition that I own a ring made of the element with atomic number 79, (2A) the proposition that there are 172 tiles on the floor, and (2B) the proposition that there are 289 tiles on the floor. Thus, Stalnaker's switch to metalinguistic contents does not properly explain the obstacle to making either advance. As the metalinguistic strategy was developed, it faced further, related problems to which it could only respond in *ad hoc* ways (see Williamson 2022 for more details, and Stalnaker 2011 and Williamson 2011 for an exchange on Stalnaker's application of his metalinguistic strategy to metaphilosophy).

A more promising strategy for the intensionalist in tracking fine–grained cognitive significance is to describe the thinker's state in terms of attitudes to a coarse–grained proposition *under a guise* (Salmón 1986 uses a similar framework, though his propositions are individuated less finely than intensions). The guise may be a sentence which expresses the proposition, perhaps combined with a context of utterance in which it does so. Thus, in advancing from (1A) to (1B), I may go from knowing the proposition they both express under the guise of the sentence (1A) to knowing the same proposition under the guise of the sentence (1B); likewise in advancing from (2A) to (2B). One truth can appear in many forms. The cognitive focus remains on the ring, and what it is made of, or on the floor, and how many tiles are on it, but through the lens of different sentences. Those advances are far from trivial.

Such an account extends smoothly to the cognitive significance of mathematical and philosophical discourse. Let S be a contentious sentence in the language of pure philosophy which in fact expresses the necessarily true proposition p. We already know p under the trivial guise of the sentence '$0 = 0$', but we do not yet know it under the non–trivial guise of the sentence S. Indeed, some sane philosophers may even accept the necessarily false contradictory of p, $\neg p$, under the non–blatantly impossible guise of the sentence not–S, even though they would of course never accept $\neg p$ under the blatantly impossible guise of the sentence '$0 = 1$'.

On this view, the intensional semantics is correct, and differences in cognitive significance between intensionally equivalent sentences are explained elsewhere in the overall account of language and thought. Sentences in

mathematical and philosophical discourse are not treated as anomalous; what makes room for their cognitive non–triviality is the same general framework that also makes room for cognitive differences between necessarily equivalent contingent sentences. The *Tractatus*' self–defeat in the philosophy of language is just one symptom of its quite general failure to provide a proper framework for understanding fine–grained differences in cognitive significance.

4 Metaphilosophy as Part of Philosophy

The examples considered in sections 2 and 3 suggest that self–defeating philosophical theories indicate the difficulty of metaphilosophy only insofar as they indicate the more general difficulty of *philosophy*. Although applying such a theory in epistemology, metaphysics, or the philosophy of language to itself is a cute dialectical shortcut, what it turns out to manifest is the inadequacy of that philosophical theory to ordinary and scientific thought and talk in general.

All this is consonant with an *anti–exceptionalist* approach to metaphilosophy: epistemologically, metaphysically, semantically, pragmatically, and in other ways, human philosophical thought and talk is not special, not radically different from human non–philosophical thought and talk. Our difficulties in making sense of human philosophical thought and talk just exemplify our difficulties in making sense of human thought and talk in general. We should be very suspicious of epistemological, metaphysical, semantic, pragmatic, or other theories that render philosophy anomalous with respect to normal human cognition, by assigning it special humiliations or, for that matter, special privileges. Those alleged anomalies are likely to illustrate ways in which the theory mischaracterizes *normal* human cognition.

This anti–exceptionalism aligns with an alternative model of how metaphilosophy is done. Instead of warping it to vindicate, or at least acknowledge, one's actual practice as a philosopher, one simply applies one's general views on epistemology, metaphysics, the philosophy of language, or whatever, to the specific case of philosophy, irrespective of any resulting discomfort. That might be a more honest way to proceed. Any resulting failure of fit will manifest some inadequacy or other in those general views.

But that alternative model may be too crude, because it implies that one's experience of doing philosophy played no role as evidence in the formation of one's general views on epistemology, metaphysics, the philosophy of language, or whatever. That would be an opportunity missed, for one's own philosophical cognition, and that of other philosophers whom one has studied, constitutes some of the cognition with which one is most familiar as a philosopher.

Reflecting philosophically on philosophical reflection, especially one's own – whether or not one thinks of it *as* philosophical – can draw one's attention to relevant cognitive phenomena that need to be understood, perhaps because they do not fit the usual stereotypes of cognition. That suggests a more positive role for metaphilosophy in philosophy, as a source of fresh data.

5 From Metaphilosophy to the Nature of Conditionals: One Philosopher's Experience

I will sketch an example from my own experience. It is a tiny fragment of the history of very recent philosophy, but it is at least one that I can claim to know as well as anyone else does. It starts with a metaphilosophical debate about the epistemology of hypothetical cases or, in more picturesque terms, *thought experiments*, and leads by a long and winding road to the general philosophical understanding of conditional thinking and the cognitive function of the imagination. It illustrates the complex, potentially fruitful interaction of metaphilosophy with both epistemology and the philosophy of language.

By the 1990s, a debate had started about the epistemology of what was called 'intuition' (see for example the papers in DePaul and Ramsey 1998). It was partly motivated by metaphilosophical anxieties, for its central prototypes of 'intuition' included unreflective verdicts on philosophers' thought experiments, like the received verdict that the protagonist of a specified Gettier case does not know. Intuitions on philosophically significant cases were called 'philosophical intuitions'. In many branches of analytic philosophy, such as epistemology and moral philosophy, philosophical intuitions were standardly treated as validating *counterexamples* to philosophical theories, for example the analysis of knowledge as justified true belief. Philosophers thereby relied on the *truth* of philosophical intuitions. Although the more holistic methodology of reflective equilibrium mandated a two–way adjustment of intuition to theory as well as theory to intuition, even that seemed to involve some initial defeasible presumption in favour of philosophical intuitions. That raised the metaphilosophical question: what reason is there to expect philosophical intuitions to be true?

The debate was soon made more urgent by the emerging 'negative program' of 'experimental philosophy'. In an influential paper, 'Normativity and Epistemic Intuitions', Jonathan Weinberg, Stephen Stich, and Shaun Nichols (2001) published survey results which seemed to show that received verdicts on some standard thought experiments in epistemology vary with ethnicity and so are unsafe. More generally, the negative program involved attempts

to assemble experimental evidence that philosophical intuitions vary with factors irrelevant to their truth, such as subjects' ethnicity and gender, and the context in which they are elicited. The proposed conclusion was that we should *not* rely on philosophical intuitions, because we have no good reason to expect them to be true – except in cases where we have experimental evidence that they are independent of those factors irrelevant to their truth. Of course, even if a philosophical intuition is humanly universal, it still does not follow that it is *true* – it might be a mere glitch in a humanly universal cognitive system. In any case, a metaphilosophical upshot of the negative program was a rejection of the case method in philosophy. The preferred alternative method of answering philosophical questions was less clear, but threatened to involve a reduction of philosophy to psychology (for a more recent general survey of experimental philosophy see Sytsma and Buckwalter (2016); my critical discussions of experimental philosophy are all reprinted in Williamson 2021; see also Nagel 2012 and Machery *et al.* 2017).

Early in the debate, both proponents and opponents of the case method characterized it as involving reliance on intuition – as many still do. When I was invited to give a talk at a workshop on intuition and epistemology at the University of Fribourg, Switzerland, in 2002, I accepted, in order to force myself to think more seriously about the issue. In my talk, I tried to fit intuition into my knowledge–first epistemology (Williamson 2000) by treating intuiting as a factive mental attitude, one which can be had only to true contents. My plan was to treat intuiting that P, like seeing that P, as a specific way of knowing that P. In apparent cases of intuiting a falsehood, one is not really in the mental state of intuiting, even if it feels just as if one were in it.

However, when I started to write up the talk more carefully for publication as part of the workshop proceedings in an issue of the journal *Dialectica*, I ran into an obstacle. I wanted some clear, simple, non–philosophical examples of the putative mental attitude of intuiting, to illustrate the proposed epistemological structure, to which I could then compare standard cases of philosophical intuiting. Postulating a distinctive kind of mental attitude exclusive to philosophizing struck me, then and now, as grossly implausible – from an evolutionary perspective, extra cognitive structure is costly, and unlikely to confer much compensating benefit if it is useful only in philosophy. But when I tried to specify the content of such paradigmatically intuitive knowledge, complexity proliferated, unlike the simplicity of much ordinary perceptual knowledge.

More specifically: Philosophers generally understood intuition and perception as mutually exclusive, by contrast with a widespread use of the terms in psychology, on which ordinary perceptual judgments count as 'intuitive' in the sense of not involving conscious reflection. Indeed, philosophers find

the epistemology of intuition puzzling *because* they take intuition to be non–perceptual. They understand intuitions as relevantly like verdicts on hypothetical cases. But articulating such verdicts on hypothetical cases takes significant linguistic complexity, if the verdicts are to be apt for truth.

For example, take a hypothetical case in moral philosophy on which you could express your verdict by saying 'I am morally obliged to jump into the pond to save the drowning child'. What is *actually* true, I presume, is not the categorical claim that you are morally obliged to jump into the pond to save the drowning child. I trust that you are not reading this chapter by the side of a pond in which a child is drowning. Rather, what is actually true is something more like the counterfactual conditional that if you *were* in the relevant hypothetical scenario, you *would be* morally obliged to jump into the pond to save the drowning child – a conditional whose antecedent characterizes the hypothetical circumstances and whose consequent is the verdict conditional on the antecedent.

Again and again, as I thought through candidate instances of intuitions, they turned out to have something like this complex conditional structure, sometimes explicit, sometimes implicit in a universal generalization. Treating the postulated special attitude of intuition as the distinctive feature of the example, while ignoring the complex structure of the sentence used to articulate the content to which the attitude was being taken, seemed quite premature. At the very least, one should take into account the contribution of linguistic structure to the distinctiveness of the examples before starting to postulate special attitudes. My paper in the workshop proceedings took a very different line from my talk at the workshop itself, and distanced itself from the postulated category of intuition; I put the word 'intuitions' in its title in ironic quotation marks (Williamson 2004).

A prototype of philosophical 'intuition' is the verdict on a Gettier case that the protagonist does not know. To adapt an example from Bertrand Russell, she forms a belief about the time by consulting a clock which, unbeknownst to her, had stopped exactly 48 hours earlier. Thanks to that coincidence, her belief that it is 3 o'clock is true. In the operative sense of 'justified', her belief is also justified, since she has no reason to suspect that the clock is not working. But, according to the standard verdict, she does not *know* that it is 3 o'clock. Thus, justified true belief is insufficient for knowledge, and the once–popular 'JTB' analysis of knowledge is mistaken. In imagining the scenario, one judges 'She does not know that it is 3 o'clock', which is just the 'offline' analogue of the 'online' judgment a third party in the scenario might make, observing people set their watches by the stopped clock.

When giving talks on philosophical thought experiments, I sometimes explained at the start of the talk that I was not using power–point because the only time I had given a power–point presentation it was a complete disaster. Later in the talk, I revealed that I had been lying: there had been no disaster, because I had never given a power–point presentation at all (I have done so subsequently, including one at the Antwerp conference). Thus, their justified belief early in the lecture that I had never given a successful power–point presentation was true, but based on a false premise, and so had not amounted to knowledge. I had made the audience experience a real–life Gettier case from the inside. The point of the exercise was to emphasize how little difference it makes evidentially whether one uses a thought experiment or a real–life experiment; working offline or online, the upshot is the same: either way, the JTB analysis is refuted. The offline aspect of thought experiments makes them seem much more distinctive cognitively than they really are.

Counterfactual conditionals in ordinary life are often easy to know: 'If I'd dropped the soup bowl, there would've been a mess', 'If you hadn't told me, I'd not have known it was painted by a ten–year old'. They often have practical or moral significance: 'If I'd turned left instead of right, I'd have got there in half the time', 'If he hadn't driven through the red light, there would've been no accident'. General scepticism about their cognitive utility is extravagant. Indeed, given the close connection between online and offline judgments, scepticism about the offline judgments infects the online judgments too. The counterfactual conditionals relevant to philosophical thought experiments are not exceptional in these respects: doubting 'If that happened to someone, they wouldn't know it was 3 o'clock' about the offline case can easily lead to doubting 'She doesn't know it is 3 o'clock' about the online case. We often learn counterfactual conditionals by making the counterfactual supposition and imaginatively drawing out its consequences in a way constrained by our background experientially–based expectations about how the world works. Philosophers' use of hypothetical cases just recruits our ordinary capacity to know counterfactual conditions. It requires no mysterious capacity of 'intuition' (see Boghossian and Williamson 2020 for a debate on the role of intuitions in philosophy).

These ideas about the epistemology of counterfactual conditionals and their application to philosophers' use of thought experiments were articulated in *The Philosophy of Philosophy* (Williamson 2007). Later, I developed them further, into a more general account of the cognitive function of the imagination (Williamson 2016). But they were provoked by the more specific debate in metaphilosophy, and my reflection on my own philosophical practice, and

the practice of many philosophers like me, in using thought experiments. This extension from metaphilosophy to general epistemology and philosophy of mind is itself an instance of anti–exceptionalism about philosophy. It works because philosophical thinking uses the same general cognitive capacities as non–philosophical thinking.

That is not the end of the story. In *The Philosophy of Philosophy*, I also applied my general account of the epistemology of counterfactual conditionals for another metaphilosophical purpose: to explain philosophers' knowledge of metaphysical modality. Kripke had identified such necessity in *Naming and Necessity* as 'necessity in the highest degree', objective, non–epistemic, real rather than nominal, as his famous examples of the necessary *a posteriori* and the contingent *a priori* demonstrated (Kripke 1972; 1980). Predictably, he faced the neo–Humean challenge to explain how we can know, given that something is actual, whether it is necessary or contingent, and given that something is non–actual, whether it is possible or impossible. My strategy was to use counterfactual conditionals as the thin end of the wedge for metaphysical modality, by showing that anyone with the cognitive capacity to handle counterfactual conditionals *already* has the cognitive capacity to handle metaphysical modalities. This was a promising line of argument because in standard combined logics for counterfactual conditionals and modal operators (read metaphysically), the modal operators are equivalent to constructions built out of counterfactual conditionals. In particular, something is impossible if and only if it counterfactually implies a contradiction, something is necessary if and only if its negation is impossible, and so on. Thus, if you can assess sentences or thoughts involving counterfactual conditionals, you can also assess sentences or thoughts involving modal operators (read metaphysically) in the same way as you assess their counterfactual conditional equivalents. This does not require you to *define* the modal operators in terms of counterfactual conditionals; you just need to treat the corresponding sentences or thoughts alike.

The equivalences are validated by standard semantic accounts of modal operators and counterfactual conditionals in a framework of possible worlds. They can also be derived by standard modal logic from two plausible and more general linking principles: first, an antecedent necessarily implies a consequent only if the former counterfactually implies the latter too; second, no possibility counterfactually implies an impossibility. So far, everything fits nicely together. Integral to this picture is the *vacuous truth of counterpossibles*: any counterfactual conditional with an impossible antecedent is true. That follows from the first linking principle by standard modal logic, for an impossibility necessarily implies anything, and so counterfactually implies anything. Semantically, no

world is a counter – instance to a counterfactual conditional whose antecedent is false at every world.

Around 2004, in the run–up to *The Philosophy of Philosophy*, I started presenting this material in talks on the epistemology of modality. One feature of the reactions slightly took me aback: the vehemence with which many able philosophers rejected the vacuous truth of counterpossibles, on the basis of what they took to be decisive counterexamples: obviously false counterpossibles. The alleged counterexamples themselves did not surprise me. *Of course*, pre–reflectively, some counterpossibles – such as 'If there were a largest prime, it would make no difference to mathematics' – are repugnant at first sight. What I found naïve or frivolous was the confidence with which carefully developed, elegant and powerful logical and semantical theories were thrown out on the basis of uncritically accepted first impressions. It only got worse when cumbersome, feeble, and unexplanatory semantic alternatives such as frameworks of impossible worlds were invoked as substitutes to vindicate whatever the first impressions happened to be (see Williamson 2024 for the methodological issues here). In teaching logic to clever students, one spends much of one's time explaining how plausible–sounding objections to standard theorems are subtly mistaken. Why think that anything more is going on when clever philosophers make plausible–sounding objections to the vacuous truth of counterpossibles?

Still, an explanatory task remains. *Why* do the 'counterexamples' seem so compelling? It is surely not just a brute psychological fact. I noticed two features of typical alleged examples of false counterpossibles. First, they are rejected on the basis of what seems to be very shallow cognitive processing: the judgment is immediate, and the antecedent's impossibility plays no apparent role. After all, they are presented as *obvious* counterexamples. Second, the opposite counterpossible is found obvious (opposite counterfactuals have the same antecedent and mutually contradictory consequents, as in 'If I were talking to the King now, my clothes would be appropriate' and 'If I were talking to the King now, my clothes would not be appropriate'). For example, we accept the counterpossible 'If there were a largest prime, it would make a difference to mathematics' without bothering to consider whether its antecedent is impossible or contingently false. We just think how much actual mathematics involves the infinity of primes. That also seems to be the basis on which we *reject* the counterpossible 'If there were a largest prime, it would make no difference to mathematics'. This suggests the hypothesis that we treat opposite counterfactuals as mutually inconsistent, so, having accepted one, we immediately reject the other.

Can opponents of the vacuous truth of counterpossibles respond that we treat opposite counterfactuals as mutually inconsistent because they *are* mutually inconsistent, so opposite counterpossibles cannot both be true? The trouble is that the inconsistency principle does not fit other parts of our practice. For, on the equally obvious–seeming principle that a conjunction counterfactually implies its conjuncts, we accept both 'If the Russell set were a member of itself and not a member of itself, it would be a member of itself' and 'If the Russell set were a member of itself and not a member of itself, it would be not a member of itself', which are opposite counterfactuals. The mutual inconsistency of opposite counterfactuals is also hard to reconcile with their natural use to articulate arguments by *reductio ad absurdum*, when we refute a mathematical hypothesis H by proving (for some R) *both* 'If H were so, R would be so' *and* 'If H were so, R would not be so'.

I therefore proposed treating the mutual inconsistency of opposite counterfactuals as a *heuristic* in the psychologists' sense: a cognitive shortcut, reliable in most but not all cases. Psychologists have studied many such heuristics intensively. Sometimes they characterize them negatively, as 'cheap and dirty', in the tradition of Daniel Kahneman (Kahneman, Slovic, and Tversky 1982), sometimes more positively, as 'fast and frugal', in the tradition of Gerd Gigerenzer (Gigerenzer, Hertwig, and Pachur 2011). At worst, heuristic–based cognition is regarded as a form of *irrationality*, at best, as a form of *bounded rationality*. Some heuristics are doubtless better than others, at least for a given purpose under given conditions. We might be better off avoiding *some* heuristics, but the nature of human cognition – perhaps of finite cognition in general – precludes avoiding them *all*. Many important heuristics are virtually universal to humans. For example, visual illusions are probably by–products of such heuristics built into the visual systems of humans and other animals (Fleming 2012, Gigerenzer 2021). We often rely on heuristics without realizing that we are doing so, often when they are built into our unconscious perceptual processing.

I argued that apparent examples of false counterpossibles are artefacts of our unconscious reliance on the heuristic for mutual inconsistency of counterfactuals. Arguably, it fails *only* for counterpossibles; since most counterfactuals are not counterpossibles, it is quite reliable. Those who take themselves to be refuting the vacuous truth of counterpossibles are arguably the victims of their own heuristics (Williamson 2017, 2018).

The heuristic that opposite counterfactuals are mutually inconsistent is easily generalized to one which treats as mutually inconsistent counterfactuals with the same antecedent and *contrary* consequents, which need not be contradictory. Thus, 'If Maria were in Europe, she would be in Italy' is treated as

inconsistent with 'If Maria were in Europe, she would be in Spain', not just with 'If Maria were in Europe, she would not be in Italy'. However, that generalization does not go very far.

Postulating a heuristic just for handling the special case of counterfactuals with the same antecedent and contrary consequents left me in an unstable position. The heuristic leads you, having accepted one counterfactual, to reject other counterfactuals related to it in the specified way, but not to accepting any counterfactuals in the first place. It is useful only if you already have some *other* way of cognitively assessing counterfactuals and sometimes accepting them. But if you have that other way, why do you need the heuristic for the special case?

Fortunately, the special heuristic hinted towards a much broader generalization. For simplicity, take the analogous case for plain conditionals without 'would'. Grant, for the sake of argument, both 'If Maria is in Europe, she is in Italy' and 'If Maria is in Europe, she is in Spain'. Suppose that Maria is Europe. On that supposition, we have both 'She is in Italy' and 'She is in Spain' about Maria, an inconsistency. By the analogous special heuristic for plain conditionals, we project that inconsistency conditional on 'Maria is in Europe' onto an unconditional inconsistency between the two original conditionals with antecedent 'Maria is in Europe'. The underlying general rule is to project a joint assessment of a set of statements on a given supposition onto the same assessment outright of the set of conditionals with that supposition as their shared antecedent and the statements in the original set as their consequents. In particular, when the original set has just one member statement, the conditional with that member statement as its consequent and the supposition as its antecedent is assessed outright as the member statement is assessed on the supposition. For example, we accept the conditional 'If Maria is in Europe, she is in Italy' outright when we accept 'She is in Italy' under the supposition 'Maria is in Europe', we reject the conditional outright when we reject 'She is in Italy' under the same supposition, and we suspend judgment on the conditional when we suspend judgment on 'She is in Italy' under the supposition.

A similar pattern applies to counterfactual conditionals, with a slightly different style of assessment appropriate to counterfactual suppositions, which are more 'distanced' from reality, by 'would', than plain conditionals. From this general heuristic, one can then recover the original principle that opposite counterfactuals are mutually inconsistent as a special case.

This large step of generalization was not meant as a compelling argument. It just explains how natural it was for me to reach the hypothesis of a simple, general, supposition–based heuristic for assessing conditionals. The hard work was then to show in detail how postulating such a heuristic explains our

assessment of conditionals in general. I did that work in my book *Suppose and Tell: The Semantics and Heuristics of Conditionals* (Williamson 2020). Once I had the central idea, writing the book went quickly, because everything fell into place much more smoothly and neatly than I had anticipated.

The suppositional heuristic was by no means a completely new idea. It is closely related to a seminal suggestion by Frank Ramsey about how we assess conditionals, known as the 'Ramsey test'. The heuristic also subsumes both the standard introduction and elimination rules for the conditional (conditional proof and *modus ponens*) in systems of natural deduction and the attractive, much–discussed, but problematic idea that the probability of a conditional should be the conditional probability of its consequent on its antecedent (see Williamson 2020 for discussion and references). It also harmonizes with the picture of the cognitive assessment of counterfactual conditionals in *The Philosophy of Philosophy*, and the central role of the imagination there. In turn, that role enables us to use conditional words like 'if' in extracting, articulating, and communicating informative connections embedded in our experientially informed ways of developing suppositions, the offline analogues of our capacities for online updating of our expectations about the future.

Attempts to build such ideas about conditionals directly into their semantics had led to disaster, as shown by various impossibility theorems proved by David Lewis and others. In brief, the general suppositional heuristic is *inconsistent* in various ways, both in itself and with uncontentious background knowledge. From my perspective, those results showed, not that the ideas should be abandoned, but that their plausibility comes from their match to our *heuristics*, not to our semantics. An inconsistent test can give correct results most of the time, but it cannot be *logically valid*.

The need for hypothetical thinking is deeply rooted in intelligent life. For example, when one has to decide between several possible actions, one must compare what would or might happen *if* one took this action with what would or might happen *if* one took another action. Thus, our primary way of assessing conditionals – the imaginative use of the suppositional heuristic – is a good candidate for a human universal. Since it is fallible, it also constitutes a potential source of error in our case judgments of conditionals. The usual methods of experimental philosophy will not pick it up, for the consequent errors may not show up in divergence between judgments made by different demographic groups or under different environmental conditions.

In postulating a fallible general heuristic for conditionals, I have encountered far more resistance from semanticists in both philosophy and linguistics than I have from psychologists and psychologically–informed philosophers.

The latter groups' reaction seems to be: 'Of course we rely on heuristics in assessing conditionals, just as we do in other cognitive tasks – how else would we do it? The question is *which* heuristics we use.' By contrast, semanticists are more likely to question the need for an intermediate level of heuristics between the semantics and our assessments of particular conditionals in particular contexts. It is as if the semantic evaluation of the conditional in the context were transparent to competent speakers and hearers; as if they could read off its semantic status directly, without cognitive effort. Although it would be uncharitable to accuse semanticists of really believing that, I often find them reluctant to admit that the cognitive task might be hard enough to require heuristics. They seem understandably afraid that admitting the role of heuristics risks losing them their data, their empirical constraints, since common native speaker assessments of sample sentences will no longer be reliable.

Such fears are over–pessimistic. The natural sciences have operated for centuries with less than fully reliable data, and have found ways to manage the risks. One such way is an aversion to what natural and social scientists call 'overfitting': profligacy with theoretical complexity, especially in multiplying the number of independent parameters in the model ('degrees of freedom'), which enables them to fit current data closely – including any erroneous data points! – but also leads to predictive failures and theoretical instability. Significantly, contemporary semantics shows increasing signs of overfitting, with very complicated semantic clauses for common words (such as 'if'), and little sense that adding new parameters of semantic evaluation might have a cost in degrees of freedom (see Williamson 2024 for more discussion). I have seen simplicity as a criterion of theory choice incredulously rejected as quite alien to semantics.

The 'paradoxes' of material (truth–functional) implication constitute a case in point. By the usual truth–table, 'If A, C' is true on the material reading of 'if' whenever A is false or C is true. It is therefore easy to manufacture apparent counterexamples to the material reading of natural language conditionals. For example, 'If my lottery ticket won, it lost' seems obviously false, but on the material reading it is almost certainly true, since its antecedent is almost certainly false. The suppositional heuristic predicts our negative judgment: we reject 'If my lottery ticket won, it lost' unconditionally because we reject 'It lost' conditionally on 'My lottery ticket won'. Nevertheless, in *Suppose and Tell*, I argue that the material interpretation of 'if' makes the best overall sense of our total practice of using conditionals to extract, store, and communicate information, and that the rejection of some true conditionals turns out to be a small price to pay for its overall utility. That account can be extended to

counterfactual conditionals, and in particular to the vacuous truth of counterpossibles. The apparent falsity of some counterpossibles is an artefact of the corresponding heuristic.

When writing the conditionals book, I naturally wondered what role heuristics might be playing in other philosophical problems. Unsurprisingly, my thoughts turned to the problem of vagueness. Many philosophers understand sorites paradoxes as generated by 'tolerance principles' for vague terms, such as 'A heap minus one grain is still a heap'. Iterating its application takes one from the obviously true 'Then thousand grains make a heap' to the obviously false 'One grain makes a heap'. Some philosophers even classify tolerance principles as 'analytic' or 'conceptual connections'. But the attraction of assigning them such semantic dignities just indicates the poverty of available options for general principles in philosophers' toolbox. Once one has the category of heuristics, a much more natural and plausible option becomes available. Tolerance principles are just labour–saving devices by which we ignore small differences and so avoid the expense in time and energy that would otherwise be spent on continually reassessing our classifications when updating on titbits of new information. To emphasize that invoking heuristics was not just an *ad hoc* move for handling conditionals, I added a few pages on tolerance principles as heuristics to *Suppose and Tell*. I even showed how to estimate their reliability in some cases: although tolerance principles generate paradoxes, in almost all instances they preserve truth. Since then, I have realized that tolerance principles are just special cases of a much more general principle, the 'persistence heuristic', on which ignoring small differences is the default: it is neither specific to vague terms nor just labour–saving, but practically indispensable for any data base.

This approach can be taken more generally to philosophical paradoxes (Williamson 2024). For example, the Liar paradox and other semantic paradoxes depend on disquotational principles for truth and falsity, on which '"Snow is white" is true' is equivalent to plain 'Snow is white' and "Snow is white" is false' to 'Snow is not white'. Such principles are needed, and almost always unproblematic, but have rare pathological instances. They can be understood as heuristics for truth and falsity. Again, Kripke's puzzle about belief can be read, against his intentions, as showing how Frege puzzles for belief like 'Hesperus'/'Phosphorus' depend on heuristics for determining what people believe from their linguistic behaviour (Kripke 1979). Heuristics for ascribing knowledge and belief can in turn cast light on the nature of knowledge and belief themselves.

In brief, the links of the chain connected like this. In response to a metaphilosophical challenge to philosophers' putative knowledge of metaphysical

modality, standard logical equivalences were invoked between metaphysical modalities and counterfactual conditionals. Those equivalences require the vacuous truth of counterpossibles, to which counterexamples were proposed – a niche dispute in philosophical logic rather than metaphilosophy. Scrutiny of the allegedly false counterpossibles indicated that what makes them look false is a form of reasoning which usually preserves truth but is untrustworthy for counterpossibles, so the alleged counterexamples were not probative. The role of that plausible but fallible form of reasoning can be explained as an application of a psychological *heuristic*. It is a special case of a general heuristic for assessing conditionals, the suppositional heuristic. The hypothesis that this heuristic is our primary way of assessing conditionals is central to an explanation of our practice of using conditionals, how we manage it cognitively, why it is so useful for us in extracting and communicating information, but also why the semantics of conditionals has baffled philosophers and logicians for over two thousand years – because our heuristic for conditionals is implicitly inconsistent, so no semantics validates all our assessments. In particular, when philosophers talk of conditional 'paradoxes', the air of paradoxicality is an artefact of their reliance on the suppositional heuristic.

That account of conditionals in turn suggests a more general hypothesis about philosophical paradoxes. They are artefacts of fallible but useful heuristics on which we unreflectively rely because they are built into our general cognitive systems, just as visual illusions are artefacts of fallible but useful heuristics on which we unreflectively rely because they are built into our visual systems. In both cases, the heuristics may be human universals.

The general hypothesis about the origin and nature of philosophical paradoxes presumably belongs to metaphilosophy. But its consequences for specific paradoxes belong to more general philosophy – topics such as conditionals, vagueness, truth and falsity, knowledge, and belief are not in themselves metaphilosophical. These interconnections exemplify the intricate relations between philosophy and metaphilosophy.

6 Conclusion

Reflecting on one's own practice can prompt one to modify that practice. Although metaphilosophy is just one small part of philosophy, it can sometimes command the rest. Through self–criticism, one may learn to do philosophy differently, perhaps even to do it better. Reflection on one's own philosophical practice can also alert one to cognitive phenomena of wider interest: one can see a world in a grain of sand. The epistemology, metaphysics, or semantics of

a philosophical statement is a sample of general epistemology, metaphysics, or semantics.

Of course, the relation of one's philosophy to one's metaphilosophy varies from philosopher to philosopher, and from tradition to tradition. It partly depends on the content of one's metaphilosophy – for example, on whether one inclines to philosophical exceptionalism or anti–exceptionalism. And some philosophers are much more prone to metaphilosophical reflection than others are. On such a matter, one should aim not to draw up general rules, but to appreciate the range of possibilities. This paper has illustrated how complex and indirect the dialectic between philosophy and metaphilosophy can be. Surely many other episodes in the history of philosophy illustrate the same point.

Acknowledgments

Thanks are due to audiences for my remote presentation of this material at the 2022 conference on 'Metaphilosophy in History' at the University of Antwerp and my live presentation of it at Yale University for helpful discussion, and to the conference organizers, Catherine Dromelet, Willem Lemmens, and Tamas Demeter, for taking the risk of inviting a metaphilosopher who is very much not a historian of philosophy to speak at such an event.

Bibliography

Boghossian, P. & Williamson, T. 2020. *Debating the A Priori*. Oxford: Oxford University Press.

Burnyeat, M. (ed.), 1983. *The Skeptical Tradition*. Berkeley: University of California Press.

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

Fleming, R. 2012. Human Perception: Visual Heuristics in the Perception of Glossiness. *Current Biology* 22, R865–R866.

Gigerenzer, G. 2021. Embodied Heuristics. *Frontiers in Psychology* 12: 711289. DOI: 10.3389/fpsyg.2021.711289.

Gigerenzer, G., Hertwig, R. & Pachur, T. (eds.), 2011. *Heuristics: The Foundations of Adaptive Behavior*. New York: Oxford University Press.

Kahneman, D., Slovic, P. & Tversky, A. (eds.), 1982. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kripke, S. 1972. Naming and Necessity. In: Davidson, D. & Harman G. (eds.), *Semantics of Natural Language*, 235–355, 763–769. Dordrecht: Reidel.

Kripke, S. 1979. A Puzzle about Belief. In: Margalit, A. (ed.), *Meaning and Use* Dordrecht: Reidel, 239–283.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell. Expansion of Kripke 1972.

Machery, E., Stich S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N. & Hashimoto, T. 2017. Gettier across cultures. *Noûs* 51: 645–664.

Nagel, J. 2012. Intuitions and experiments: a defense of the case method in epistemology, *Philosophy and Phenomenological Research*, 85: 495–527.

Quine, W. O. 1951. Two Dogmas of Empiricism. *Philosophical Review* 60, 20–43.

Salmón, N. 1986. *Freges Puzzle*. Cambridge, Massachusetts: MIT Press.

Stalnaker, R. 1984. *Inquiry*. Cambridge, Mass: MIT Press.

Stalnaker, R. 1999. *Context and Content*. Oxford: Oxford University Press.

Stalnaker, R. 2011. The Metaphysical Conception of Analyticity. *Philosophy and Phenomenological Research* 82, 507–514.

Sytsma, J. & Buckwalter, W. (eds.), 2016. *A Companion to Experimental Philosophy*, 22–36. Oxford: Wiley–Blackwell.

Unger, P. 1979. I do not exist. In: Macdonald, G. (ed.), *Perception and Identity*. London Palgrave, 235–251.

Weinberg, J., Stich, S. & Nichols, S. 2001. Normativity and Epistemic Intuitions. *Philosophical Topics* 29, 429–460.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Williamson, T. 2004. Philosophical "Intuitions" and Scepticism about Judgment. *Dialectica* 58, 109–153.

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

Williamson, T. 2011. Reply to Stalnaker. *Philosophy and Phenomenological Research* 82, 515–523.

Williamson, T. 2016. Knowing by Imagining. In: Kind, A. & Kung, P. (eds.), *Knowledge through Imagination*, 113–123. Oxford: Oxford University Press. Reprinted in Boghossian, P. & Williamson, T. 2020. *Debating the A Priori*. Oxford: Oxford University Press, 175–185.

Williamson, T. 2017. Counterpossibles in semantics and metaphysics. *Argumenta* 4. URL: https://www.argumenta.org/article/counterpossiblea-semantics-metaphysics/.

Williamson, T. 2018. Counterpossibles. *Topoi* 37, 357–368.

Williamson, T. 2020. *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford: Oxford University Press.

Williamson, T. 2021. *The Philosophy of Philosophy*, enlarged edition. Oxford: Wiley–Blackwell.

Williamson, T. 2022. Metametaphysics and Semantics. *Metaphilosophy* 53, 162–175.

Williamson, T. 2024. *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.