sciendo

Disputatio's 25th Anniversary Lecture

Degrees of Freedom: Is Good Philosophy Bad Science?

Timothy Williamson
University of Oxford

DOI: 10.2478/disp-2021-0005 BIBLID [0873-626X (2021) 61; pp.73–94]

Abstract
The lecture starts by considering analytic philosophy as a tradition, and its global spread over recent years, of which *Disputatio*'s success is itself evidence. The costs and benefits of the role of English as the international language of analytic philosophy are briefly assessed. The spread of analytic philosophy is welcomed as the best hope for *scientific* philosophy, in a sense of 'science' on which mathematics, history, and philosophy can all count as sciences, though not as natural sciences. Arguably, experimental philosophy provides no plausible alternative methodology for philosophy, only a way of psychologizing it. However, it serves a useful purpose by highlighting the inadequacy of current methods for detecting errors in judgments on possible cases, which may result from reliance on possibly universal but imperfectly reliable cognitive heuristics. The problem is exacerbated by analytic philosophers' tendency to regard increased flexibility in a theoretical framework as progress, where natural scientists would treat it as methodologically vicious profligacy with degrees of freedom. The result is a familiar type of bad science, *overfitting* theory to uncritically accepted data. The recent 'hyperintensional revolution' may be an example of such overfitting, it is suggested. The lecture ends with a call for a more miserly attitude to degrees of freedom.

Keywords
Analytic philosophy, tradition, hyperintensional, heuristic, sorites.

This lecture celebrates an historic occasion: the 25th anniversary of the journal *Disputatio*. I will start with a few historical remarks, sketching a big picture without attempting to fill in the details, and then move on to issues in contemporary philosophy.

When we look back on the history of philosophy in the late

twentieth and early twenty-first centuries, one of the most striking developments has been the global spread of *analytic philosophy*, understood not in any narrowly doctrinal or methodological sense, but as a broad, dynamic tradition, always in flux, never homogeneous, yet held together by a complex network of interacting influences. In that way it is like many other intellectual traditions flexible enough to be capable of continually renewing itself. Such an understanding of the phrase 'analytic philosophy' best fits the way it is actually applied. In that spirit, *Disputatio* defines its remit as 'all aspects of analytical philosophy (broadly construed)'. Indeed, the founding and flourishing of *Disputatio* is part of the global development just indicated.

The journal's title alludes to the institution of *disputatio* in medieval universities, debates structured by formal rules, motivated in turn by principles of logic. Like the Petrus Hispanus Lecture series at the University of Lisbon, it rightly affirms the underlying connections between analytic philosophy and much older traditions of rigorous philosophy.

Nevertheless, the phrase 'analytic philosophy' is most useful when confined to a historically delimited tradition, rooted in the late nineteenth century, and associated with the rise of modern logic. Much of the early action was in central Europe, but the political catastrophes of the mid-twentieth century largely expunged the tradition from those original homelands, while the consequent flow of refugees spread it elsewhere, above all to the United States. As a result, for decades after 1945, analytic philosophy was almost (but not quite) exclusively confined to the English-speaking world, in the sense of those regions where English is the dominant native language. Of course, such monolingualism was always anomalous; over my career, I have watched its gradual disappearance with relief. Analytic philosophy is now practised all over the globe, whatever the native language of the region—for example, Portuguese.

But doesn't English *remain* the dominant language of analytic philosophy? If you want to influence the direction of research on just about any given topic within it, you had better publish in English. Thus, although Portuguese and English are the two official languages of *Disputatio*, in practice almost all its articles are in English. However, the reason *why* English dominates analytic philosophy is quite different from before. Analytic philosophy is now extensively practised

in non-English-speaking countries, but English remains its dominant language for the same reason as it is the international language of science, just as Latin was the European language of learning in the middle ages. If you want to make a difference to the direction of research on just about any given topic in natural science, you had better publish in English.

Of course, the role of English as the international language of science has to do with the economic, political, and cultural position of the United States in the post-war period, and it gives an unfair head start to native speakers of English. But it does not prevent non-English speaking countries from producing world-class research in natural science. Science benefits enormously from having a universal language of communication, and for the present English is the only feasible candidate. Similar remarks apply to philosophy, especially when conceived as part of the scientific enterprise, as much analytic philosophy has historically been, not least in the spirit of logical positivism.

Admittedly, some analytic philosophers reject any such 'scientific' conception of philosophy. The most salient example is Wittgenstein, who nevertheless clearly belongs in the analytic tradition by the operative standard of his place in the relevant network of influences. His heirs, and others with a more 'literary' conception of philosophical writing, may feel more pull to exploit the full resources of their native language, and less to write in the international language of science. For similar reasons, 'ordinary language philosophy' done in English tends to marginalize non-fluent speakers of English.

Even such a scientific sub-discipline as the formal semantics of natural language can work in a similar way when the examples are mainly drawn from English. Although it is well understood that in principle all natural languages are equally good sources of examples, implementing that attitude in practice is harder: a robust methodology requires sample sentences to be evaluated by several native speakers, not just one, making a quorum harder to attain. Still, the philosophy of language has much to gain from a more multilingual approach, since it should help us avoid errors of over-generalization from special features of English, or any other single language.

We can also consider more abstractly the outward spread of an intellectual tradition from an initial base. At least in the early stages

of the process, we can expect that base to remain the *centre* of the tradition—for instance, analytic philosophy—with a growing periphery outside. The asymmetry between centre and periphery manifests itself in terms of the balance of intellectual trade: the centre is a net exporter of ideas, the periphery a net importer. Bibliometrics would provide a crude measure of the imbalance: the periphery cites the centre more than the centre cites the periphery. Another mechanism is travel to the initial base for purposes of study, and then return. If I may be permitted the vanity of a personal example, João Branquinho—who took the lead in the founding of *Disputatio*—came to Oxford in 1988 for his doctoral studies, and was my first DPhil student there, before returning to be a Professor at the University of Lisbon. I must admit, though, that I had very little to teach him about rigour—he already had it in a high degree when he started.

In the longer run, however, the tradition has no need to remain as centralized as in those early stages. The net balance of trade in ideas may even be reversed. In particular, analytic philosophy since 1945 has become less and less centralized, first through the multiplication of 'centres' within the English-speaking world and then through the spread of new 'centres' far beyond that world, where each centre produces ideas for export. A salient example is of course the Language, Mind and Cognition Group (LanCog) in Lisbon, with which *Disputatio* is associated. Such a distributed tradition is healthier, for its larger gene pool of ideas makes it more robust. There is less danger of inbreeding.

Predictably, I regard the spread of analytic philosophy as *progress*. However, in characterizing it as an intellectual tradition, I did not define analytic philosophy by its virtues, but only by a particular network of influences. For all I said, those could be *bad* influences. Plainly, not all contemporary analytic philosophy is good philosophy, and not all contemporary *non*-analytic philosophy is bad philosophy. Moreover, the analytic tradition is *dynamic*, always in flux. In the mid-twentieth century, its most prominent strands were logical positivism and ordinary language philosophy, both of which arguably made very significant contributions to philosophical progress in their time, but nevertheless depended on now-obsolescent conceptions of linguistic rules. If the present global spread of analytic philosophy were merely the spread of logical positivism or ordinary language

philosophy, I would see little to celebrate in that spread. Fortunately, analytic philosophy, having learnt from both movements, has moved on.

So why *is* the global spread of the analytic tradition something to celebrate? In short, because for the foreseeable future the analytic tradition is the main home of *scientific philosophy*.

I do not use the term 'science' in the narrow sense of natural science, which involves the use of experimental and observational methods. For instance, mathematics is a science, but not a *natural* science. I use 'science' in the much broader sense of systematic, critical, evidence-constrained inquiry into how things are on some topic. In that sense, the rigorous study of *history* based mainly on written documents also counts as a science, though not as a natural science. Thus scientific philosophy is philosophy done as a science, but that does not require the assimilation of philosophy to a *natural* science. Where appropriate, scientific philosophy will use the methods of modern logic and mathematics, or those of history. The evidence which constrains it can be anything we know. Where relevant, it includes whatever we know from natural science, but also from logic and mathematics and from history, and whatever else we know by using our usual cognitive faculties (there is no need to postulate a distinctive class of 'philosophical intuitions'). Scientific philosophy follows the same general abductive standards for theory choice as the rest of science. Since clarity, precision, and rigour tend to serve those standards, scientific philosophy favours those qualities as far as feasible, again like the rest of science.

Of course, any such conception is contentious. Many philosophers regard scientific philosophy in that sense as impossible; some regard it as undesirable *even if* possible. I have defended both the possibility and the desirability of scientific philosophy elsewhere, and will not repeat that defence here. I am just putting my cards on the table.

I suggest that many—perhaps most—great philosophers before the analytic tradition did something like scientific philosophy, and that much—though not all—analytic philosophy, including, I am sure, all contributions to *Disputatio*, approximates to scientific philosophy too. Presumably, most hard-line Wittgensteinians will deny that they are doing anything like scientific philosophy, and we can provisionally take them at their word. Conversely, I do not assert

that *all* contemporary scientific philosophy is analytic philosophy. Some textually and philologically meticulous research on the history of philosophy may count as scientific philosophy, given my broad understanding of 'science', even though it is not involved in the network of influences which constitutes the analytic tradition. However, *in my experience*, most contemporary research on non-historical philosophical questions with a serious claim to be scientific philosophy *does* belong to the analytic tradition, and the spread of that tradition represents by far the best hope for the growth of scientific philosophy in the contemporary world.

Still, congratulating ourselves on how scientific we are is not the best use of analytic philosophers' time. We do better to think about how we can become *more* scientific. That is my agenda for the rest of this lecture. I will explain a specific way in which analytic philosophy is not yet scientific *enough*.

I will start with probably the most scientific branch of philosophy: logic. In debates on alternative logics, a common move is to identify a putative principle of logic (such as the Law of Excluded Middle) whose failure one can in some sense conceive, and to assert: even if this principle in fact holds, it is too substantive to count as *logically* valid; it forecloses a question which logic should leave open. This move can *sound* admirably open-minded, and in that way scientific. Call the underlying conception *Neutral Logic*.

Obviously, the general tendency of Neutral Logic is to weaken logic, to make as little as possible count as logically valid. Indeed, with sufficient ingenuity, one can in some sense conceive the failure of *any* putative principle of logic. Thus the natural conclusion of Neutral Logic is that *no* principle is logically valid, so logic is entirely useless. People who rely on Neutral Logic in arguing against classical logic typically conceal this natural consequence of their starting-point from themselves, and perhaps from others, by focussing myopically on just one way of conceiving the failure of a putative principle. As a fall-back, they might claim that a small dose of Neutral Logic is healthy, even if gorging on it would be fatal.

Consider the analogue of a small dose of Neutral Logic for mathematics as applied in natural science. One rejects a specific principle of standard mathematics because one can in some sense conceive its failure, claiming that it thereby forecloses a question which

mathematics should leave open. Such a move would reduce the explanatory power of theories in natural science by eliminating some of their mathematical consequences. Of course, in particular applications one could still recover the relevant consequences by postulating them as *ad hoc* auxiliary hypotheses, but postulating an auxiliary hypothesis instead of deriving it from independent general principles still reduces explanatory power. The same goes for applying a small dose of Neutral Logic to the background logic in natural science. Not even a small dose of Neutral Logic is good science. It is a refusal to come to terms with the inherently risky nature of the enterprise of scientific thinking.

Logic is not a special case. Similar issues arise for any theoretical framework. There is a trade-off between flexibility and inferential power. The more flexible the framework, the less powerful it is in extracting consequences from hypotheses. Conversely, the more powerful it is in extracting consequences from hypotheses, the less flexible it is. This trade-off is familiar in science. The flexibility of a model can be roughly measured by the number of its *degrees of freedom*, of adjustable parameters in the model. One might expect that the more degrees of freedom, the better. Isn't freedom a good thing? The more parameters to be adjusted, the greater the model's capacity to handle whatever nature throws at it. But that is not how scientists think. They regard a large number of degrees of freedom as a *vice*, not a virtue. By adding more and more degrees of freedom, one can fit just about any data, but in a cheap way which typically brings no insight.

The problem with having too many degrees of freedom is not just uninformativeness. It is also insensitivity to errors in the data, since the model can accommodate any data, however anomalous. An outlying data point stands out as a large potential error if you restrict yourself to fitting the data with a linear function, which has few degrees of freedom. By contrast, if you allow yourself unlimited degrees of freedom, you can always fit the outlying data point with a polynomial function of sufficiently high degree. But the latter approach typically leads to severe instability in the selected values of the adjustable parameters, and so in one's hypotheses. To accommodate new data, you may need a polynomial of even higher degree, with quite different parameter values. The process of adjusting the

parameters fails to converge. Scientists call that methodological vice *overfitting*. Too much freedom is bad science.

Of course, not all functions in nature *are* linear. A quantity you are interested in may genuinely be related to another by a polynomial function of higher degree. But you are unlikely to find that function by adding more degrees of freedom whenever you need them for an exact fit with the data. You are far more likely to be sent off on a wild goose chase by excessive respect for data which contain some inaccuracies. A better methodology is to be very reluctant to add new degrees of freedom, doing it only after potential sources of error in the data have been properly investigated and shown not to explain the evidence just as well.

Do these methodological morals apply to analytic philosophy? They should make us think more carefully about the standard method of *conjectures and refutations*. It is in effect the analogue for philosophy of a rather naïve version of Karl Popper's falsificationism for natural science. Universal generalizations are conjecturally proposed, and tested by the search for falsifying instances. If such a counterexample is found, the conjecture is refuted, and the cycle begins again with a new conjecture. Whereas counterexamples in natural science come from real-life experiments and observations, in philosophy they often come from *thought experiments*. The logical basis of falsificationism is the elementary point that even one counterexample is inconsistent with the universal generalization. Epistemologically, however, the situation is more complex. How do we ensure that the supposed falsifying instance is not based on some sort of error? The danger is that we then dismiss a *true* universal generalization as refuted once and for all, and never return to it.

One standard precaution in natural science is to require experiments to be *repeatable*. Thought experiments in philosophy are *de facto* subject to a similar requirement, since the community will not accept them if other philosophers do not share the original verdict. But repeatability does not suffice against all potential sources of error. For example, suppose that the result of the relevant type of experiment depends on a systematic interfering factor of which the scientific community is unaware. The experiment has the same result each time it is repeated, but the interpretation of the result as establishing a falsifying instance of the conjecture at issue depends on neglecting

the interference, which happens each time. Sometimes errors are repeatable too. In such cases, even though the experiment is repeatable, simply to take the alleged falsifying instance at face value would lay one open to the charge of overfitting. A wider-ranging search for potential sources of error is more appropriate.

Proponents of the 'negative program' of experimental philosophy have campaigned against taking *any* results of philosophical thought experiments at face value. Much of the original campaign targeted repeatability. Some early results suggested that although the standard verdicts on classic philosophical thought experiments were repeatable amongst white males, they were not repeatable amongst non-whites and non-males, and so should not be treated as reliable. However, most of those early results were in turn later found not to be repeatable themselves, once the experiments were redone in accord with the standards prevalent in experimental psychology. Some framing effects may be more robust, such as some variations in verdict correlated with the order in which thought experiments are presented, but they are not very relevant to the long-term impact of thought experiments on the philosophical community, where verdicts must retain conviction across a wide range of framing contexts.

In general, the negative program has not lived up to its early promise, or threat. As a program for the reform of philosophical methodology, it always had the central weakness that it offered no credible alternative.

Of course, there was the idea that everyone should do experimental philosophy, but it faces the problem that experimental philosophy does not address most philosophical questions. For example, consider the question 'Is torture always wrong?' Experimental philosophy may be able to find out whether *most people think* that torture is always wrong, but that is not the same question as whether torture *is* always wrong. We could link the two questions given the (not especially plausible) bridge principle that torture is always wrong if and only if most people think that torture is always wrong. But experimental philosophy cannot establish that bridge principle. At best, it can establish whether *most people think* that the bridge principle holds, but that is not the same question as whether the bridge principle *does* hold. We could link the two questions given the (not especially plausible) bridge meta-principle that the bridge principle holds if and

only if most people think that that the bridge principle holds, but that is to embark on an infinite regress which will never pull any of the required links out of the hat. To insist that philosophers should do *only* experimental philosophy is in effect to tell them to stop asking philosophical questions and do psychology instead.

In practice, proponents of the negative program made only an apparently less extreme demand: that philosophers should stop relying on 'philosophical intuitions', such as verdicts on philosophical thought experiments. However, they were unable to demarcate a distinctive psychological category of 'philosophical intuitions' narrower than the category of all judgments not based on conscious reflection. But all reflective judgments ultimately depend on unreflective judgments! Thus their methodological injunction inadvertently implied a ban on relying on any of our judgments at all, and so committed experimental philosophers to a general scepticism very far from their intentions. In response, some experimental philosophers tried restricting the ban to unreflective judgments about 'unusual' situations, such as those postulated in most philosophical thought experiments, on the grounds that our unreflective judgments had not been calibrated to deal with such situations. However, the class of 'unusual' situations remained hopelessly ill-defined. After all, linguists since Chomsky have emphasized that our language faculty must be able to deal properly with utterances of sentences we have never previously encountered. Indeed, we are *always* in an unusual situation, once enough detail is included. Thus a ban restricted to unreflective judgments about 'unusual' situations in effect licences experimental philosophers to apply their ban *ad hoc*, however it happens to suit them. That is hardly good science.

In short, the negative program of experimental philosophy offers no coherent alternative methodology consonant with its extreme rhetoric. Nevertheless, its radical gestures do point towards a genuine weakness in the current methodology. Philosophers are surely *fallible* in their verdicts on thought experiments—not because thought experiments are more problematic in principle than other forms of argument in philosophy, but because philosophers are human. For reasons already explained, the kind of fallibility we need to worry about is not idiosyncratic aberration but something much more systematic. In particular, we should worry about potential

sources of error built into the normal human cognitive apparatus, so that the errors they produce may even be human universals. As a result, a false verdict on a thought experiment may achieve consensus across ethnicities and genders. The usual methods of experimental philosophy are ill-designed to pick up that kind of unreliability. But the more urgent worry is that the standard methods of *analytic* philosophy are also ill-designed to pick it up, because they tend towards overfitting. When there is consensus amongst philosophers as to the natural verdict on a thought experiment, they tend to accept that verdict as correct without more ado, and adjust theory accordingly.

Of course, the possibility in principle of humanly universal error is in itself just one more sceptical scenario, and no good argument for a sceptical conclusion. But in some cases we have positive reason to postulate such errors. In particular, we may have psychological evidence that the normal human cognitive apparatus includes various *heuristics*, 'fast and frugal' or 'quick and dirty' ways answering questions which are reliable enough to be useful, but still not perfectly reliable. For example, susceptibility to various visual illusions is built into the normal human apparatus for perceptual judgment, as a by-product of otherwise valuable heuristics. The consequent errors are humanly universal just in the sense that humans are normally predisposed to make them, though we can inhibit that disposition to error. In the Müller-Lyer illusion, we are tempted to judge that one line is longer than the other, but, once we are in the know, we refuse to give way to temptation. Heuristics are not limited to perception; we need and use them in other cognitive processes too. Often, we have no pre-theoretic access to the fact that we are using an imperfectly reliable heuristic, or to what heuristic it is, or to its status as a mere heuristic rather than a universal rule. Such heuristics, I suggest, can produce cognitive illusions which lead us to make false judgments in philosophically relevant cases. I will propose some examples.

In his paper 'A Puzzle about Belief' (1979), Saul Kripke explains how our normal ways of ascribing beliefs can easily lead us into inconsistency in quite ordinary cases. He plausibly suggests that English speakers rely on something like the schema "*A normal English speaker who is not reticent will be disposed to sincere reflective assent to 'p' if and only if he believes that p*". Plausibly, users of other natural languages rely on analogous schemata. Call this family of schemata *the assent principle*.

Combined with the convincing principle that correctly translating a belief-ascription preserves its truth-value, the assent principle generates inconsistency in describing the beliefs of a bilingual speaker in realistically possible circumstances (the famous case of puzzling Pierre). Indeed, as Kripke also explains, the problem arises even in the monolingual case.

Readers of 'A Puzzle about Belief' may interpret Kripke as suggesting that our ordinary concept of belief is *incoherent*, or something like that. In response, some may attempt to qualify the assent principle in more or less elaborate ways to avoid the contradiction. But there is a simpler possibility. The assent schema may be a normal *heuristic* for ascribing beliefs. After all, it provides a quasi-operational test for belief in terms of overt behaviour. Of course, the words 'normal', 'reticent', 'disposed', 'sincere', and 'reflective' provide some wiggle room, presumably intended to avoid obvious counterexamples. But the underlying heuristic could be even simpler: "*An English speaker will assent to 'p' if and only if they believe that p*". All the complications may be fall-backs we invoke when the basic heuristic fails. If this approach is correct, many of the apparent counterexamples to various theoretical claims presented in the voluminous and inconclusive literature on propositional attitude ascriptions may be errors generated by our implicit reliance on fallible heuristics. The many complicated accounts proposed for the semantics of propositional attitude ascriptions may just be artefacts of overfitting.

Another potential case of overfitting is the semantics of conditionals. In my book *Suppose and Tell: The Semantics and Heuristics of Conditionals* (2020), I argue that our primary heuristic for assessing conditional sentences is the *Suppositional Rule*, closely related to what is often called the *Ramsey Test*, after Frank Ramsey. The rule is this: your hypothetical assessment of 'C' on the supposition 'A' should be the same as your non-hypothetical unconditional assessment of 'If A, C' (also when both assessments are made on some auxiliary background suppositions). Thus if you hypothetically accept 'C' on the supposition 'A', you should non-hypothetically accept 'If A, C'. Similarly, if you hypothetically reject 'C' on the supposition 'A', you should non-hypothetically reject 'If A, C'. Moreover, your probability for 'C' conditional on 'A' should equal your unconditional probability for 'If A, C'. This rule fits how we naturally assess conditionals

by imagining that the antecedent holds and asking ourselves on that supposition whether the consequent holds. Nevertheless, the Suppositional Rule cannot be perfectly correct, for in many circumstances it mandates mutually inconsistent assessments, as can quite easily be shown (I omit details). Thus the Rule is best understood as a less than fully reliable heuristic whose distinctive cognitive function is to extract information in sentential form from our capacity to imagine in reality-oriented ways. But the literature on conditionals in natural language relies extensively on assessments of sample conditional sentences in line with the Suppositional Rule. Such assessments have been taken to motivate increasingly complicated semantic theories of plain conditionals, which freely help themselves to additional degrees of freedom (starting with a parameter for a similarity relation or selection function). In particular, such data are taken to refute the maximally simple material (truth-functional) interpretation of 'if'; that interpretation adds no degree of freedom to the semantics at all. Thus most current semantic theories of natural language conditionals add degrees of freedom to accommodate data sets arguably generated by an inconsistent heuristic. This looks like a classic case of overfitting.

Heuristics may also underlie traditional philosophical paradoxes. For example, the Liar paradox results from a disquotational principle for truth, which is central to the ordinary practice of assessing truth-ascriptions. Nevertheless, we might understand the disquotational principle as a heuristic which is reliable in all normal cases.

Similarly, the Sorites Paradox is often attributed to a tolerance principle like this: if two arrangements of grains differ only in the presence of one grain, then one arrangement constitutes a heap if and only if the other does. On most views of vagueness, such a tolerance principle cannot be perfectly true, but is still 'almost true'. In particular, on an epistemicist view, although it has false instances, *almost* all its instances are true. Thus it is a natural candidate for a heuristic. It conveniently economizes on judgment: once we have determined perceptually whether an arrangement constitutes a heap, the tolerance principle permits us to continue relying on the resulting classification under slight variations of the arrangements, without wasting time on a new perceptual determination (after all, the arrangement might no longer be in sight).

If the Liar and Sorites paradoxes depend on erroneously treating handy heuristics as would-be universal semantic requirements, then strategies of invoking non-classical logic or non-bivalent semantics—in effect, helping oneself to more degrees of freedom—look like massive overreactions: yet more cases of overfitting.

In the final part of this lecture, I will discuss in more detail another potential case of overfitting. First, some background. One major change in analytic philosophy is now often labelled 'the possible worlds revolution' or 'the intensional revolution'. It is especially associated with the work of Saul Kripke, Robert Stalnaker, and David Lewis, though its antecedents can be traced further back in the history of modal and tense logic. In the 1960s and early 1970s, the main ideas of possible worlds semantics for modal logic became an important part of the intellectual toolkit of a younger generation of analytic philosophers, to be used wherever they might help. The semantic framework was 'intensional' rather than 'extensional' because it treated modal operators as operating on the intension of a sentence—its profile of truth-values across worlds—rather than on its extension—its truth-value at the actual world. Such frameworks also became standard in formal semantics as a branch of linguistics.

Intensionalists regard extensional distinctions as typically too coarse-grained to do the required work. But then, from the 1980s on, a new wave of analytic metaphysicians started regarding intensional distinctions as *themselves* too coarse-grained to do the required work. An early example was the complaint against using the modal idea of *supervenience* to capture the relation between mental properties and physical properties (no possible mental difference without a possible physical difference), that since supervenience is not an asymmetric relation, it does not capture the asymmetric way in which physical properties are supposed to *determine* mental properties. A turning-point was Kit Fine's seminal critique in the early 1990s of Kripke's modal rehabilitation of essential properties as necessary properties of an individual, which had been taken as one of the explanatory successes of the intensional revolution. Increasingly, metaphysicians started to hold that intensionality is not where the action is in metaphysics, that it is instead with something more fine-grained, such as determination, grounding, fundamentality, or whatever. People now talk of the 'hyperintensional revolution' as having superseded the

intensional revolution.

Methodologically, a key feature of the hyperintensional revolution is that it is driven by *examples*, especially by apparent counterexamples to intensional principles. The main argument of Fine's classic paper 'Essence and Modality' (1994) is based on contrasts like that between (1A) and (1B):

(1A) It is essential to Socrates that he is Socrates.

(1B) It is essential to Socrates that he is a member of {Socrates}.

If one is willing to think in essentialist terms at all, one is liable to be struck by (1A) as true and by (1B) as false, at least until conscious theoretical commitments intervene; (1A) sounds like a truism, (1B) like the introduction of something extraneous to Socrates himself, his singleton set. Nevertheless, on standard views of the modal metaphysics of sets, being Socrates is necessarily equivalent to being a member of {Socrates}. Thus, if (1A) and (1B) differ in truth-value, the sentence operator 'It is essential to Socrates that' is hyperintensional. Consequently, essentialist ideas cannot be understood in purely intensional terms.

Fine's argument is powerful. It is valid in form, and his assessments of pairs such as (1A) and (1B) have been widely found convincing, not only by people with prior sympathy for his conclusion. Let us consider some more arguments of similar form for hyperintensionalist conclusions about other matters.

Here is an argument that 'because' (read constitutively) is hyperintensional. Consider (2A) and (2B) (assume that grass *is* green):

(2A) The proposition that grass is green is true because grass is green.

(2B) Grass is green because the proposition that grass is green is true.

It is natural to assess (2A) as true and (2B) as false (a contrast anticipated by Aristotle): truth comes from the world, not *vice versa*. Nevertheless, on standard views of propositions and truth, 'The proposition that grass is green is true' is necessarily equivalent to 'Grass is green' (importantly, sentential disquotation is not involved). There

is no intensional difference between (2A) and (2B). Thus 'because' is hyperintensional, the argument concludes.

We can reach the same conclusion by using a repetitious conjunction in place of a truth-ascription:

(3A) Grass is green and grass is green because grass is green.

(3B) Grass is green because grass is green and grass is green.

It is natural to assess (3A) as true and (3B) as false: we no more need 'and' than we need 'true' to express the underlying fact. Nevertheless, 'Grass is green and grass is green' is necessarily equivalent to 'Grass is green'.

Here is a third argument of the same form for the hyperintensionality of 'because'. Let Vera be a vixen (a female fox):

(4A) Vera is a vixen because Vera is a female fox.

(4B) Vera is a female fox because Vera is a vixen.

It is natural to assess (4A) as true and (4B) as false: we do not need 'vixen' to express the underlying fact. Nevertheless, 'Vera is a vixen' is necessarily equivalent to 'Vera is a female fox'.

The trouble is that 'vixen' is *synonymous* with 'female fox', so by compositional semantics 'Vera is a vixen' is synonymous with 'Vera is a female fox' and (4A) is synonymous with (4B). But if (4A) and (4B) are synonymous, they cannot differ in truth-value! Something has gone wrong with the argument.

One hypothesis is that the contrasting assessments of (4A) and (4B) were correct but 'because' covertly creates a quotational context in which one is talking about the linguistic expressions 'vixen' and 'female fox', not about the non-linguistic property of being a vixen, in other words, the non-linguistic property of being a female fox. But if 'because' can covertly work like that in (4A/B), we need to determine whether it is working like that in (2A/B) and (3A/B) too.

That is not what hyperintensionalists typically want. To make their hyperintensional point, they need to use 'because' with a *non*-metalinguistic, metaphysical reading. On that reading, (4A) and (4B) cannot differ in truth-value, so one of the original assessments was

incorrect. Presumably, we made the error because we were distracted by superficial linguistic features of the examples. But since (2A/B) and (3A/B) are so similar to (4A/B), we need to check whether we made similar errors in assessing the former, again distracted by superficial linguistic features of the examples. The answer is by no means obvious.

Let us consider another case, this time involving the causal construction 'brought it about that'. Here is an argument that it is hyperintensional in its second argument:

(5A) Mary brought it about that John was a contributor.

(5B) Mary brought it about that John was a self-identical contributor.

Imagine an ordinary context in which (5A) is true. It is tempting to assess (5B) as false, on the grounds that Mary did not bring it about that John was self-identical: his self-identity had nothing to do with her. Nevertheless, 'John was a self-identical contributor' is necessarily equivalent to 'John was a contributor'. Thus 'Mary brought it about that' is hyperintensional, the argument concludes.

Here is a similar example. It concerns events in 1461, during the English War of the Roses. Richard is Richard Neville, Duke of Warwick, known as Warwick the Kingmaker; Edward is Edward Plantagent, who became King Edward IV.

(6A) Richard brought it about that Edward was a king.

(6B) Richard brought it about that Edward was a male monarch.

As a matter of historical fact, (6A) is true. It is tempting to assess (6B) as false, on the grounds that Richard did not bring it about that Edward was male: his maleness had nothing to do with Richard. Nevertheless, 'Edward was a male monarch' is necessarily equivalent to 'Edward was a king'. Thus 'Richard brought it about that' is hyperintensional, the argument concludes.

The trouble is that 'king' is *synonymous* with 'male monarch', so by compositional semantics 'Edward was a king' is synonymous with 'Edward was a male monarch' and (6A) is synonymous with (6B). Thus (6A) and (6B) cannot differ in truth-value. Since (6A) is a

well-known historical truth, (6B) is true too. Thus, although Richard did not bring it about that Edward was male, it does not follow that he did not bring it about that Edward was a male king. But that has implications for our assessment of (5B) too: although Mary did not bring it about that John was self-identical, that does not show that she did not bring it about that he was a self-identical contributor.

What went wrong when we first assessed (6B)? In the example, 'brought it about' is used in a robustly causal sense; a covertly quotational reading is not plausible. More likely, the explicit inclusion of the otherwise redundant word 'male' somehow misled us into envisaging Richard's having brought it about that Edward was male as an extra condition for the truth of (6B). Similarly, perhaps, the explicit inclusion of the otherwise redundant phrase 'self-identical' somehow misled us into envisaging Mary's having brought it about that John was self-identical as an extra condition for the truth of (5B). In such ways, superficial linguistic features can easily deceive us into accepting unsound arguments for hyperintensionality.

Something like that may be happening with (2A/B), (3A/B), and (4A/B). In particular, suppose (plausibly) that we tend to assess the truth-value of statements of the form 'Y because X' by assessing the helpfulness of the corresponding (putative) explanations of the form 'X. So Y'. Thus the statements (2A/B), (3A/B), and (4A/B) correspond to the (constitutive) explanations (2A*/B*), (3A*/B*), and (4A*/B*) respectively:

(2A*) Grass is green. So the proposition that grass is green is true.

(2B*) The proposition that grass is green is true. So grass is green.

(3A*) Grass is green. So grass is green and grass is green.

(3B*) Grass is green and grass is green. So grass is green.

(4A*) Vera is a female fox. So Vera is a vixen.

(4B*) Vera is a vixen. So Vera is a female fox.

As explanations, (2A*), (3A*), and (4A*) look better than (2B*), (3B*), and (4B*) respectively. In both (2A*) and (3A*), the direction

of explanation is from the simpler and more elementary to the more complex and less elementary, which (other things being equal) is the cognitively helpful direction, and so the right one. By contrast, in (2B*) and (3B*) the direction of explanation is the opposite, which (other things being equal) is the cognitively unhelpful direction, and so the wrong one. In (4A*), the movement is from two words to one, but the two words ('female' and 'fox') seem simpler and more elementary than the one ('vixen'), so the direction of explanation is still the cognitively helpful one, whereas in (4B*) it is the opposite. Thus the helpfulness of the explanations (2A*/B*), (3A*/B*), and (4A*/B*) predicts the perceived truth-value of the corresponding 'because' statements (2A/B), (3A/B), and (4A/B).

The crucial point is this. The helpfulness of a (putative) explanation is sensitive to its superficial linguistic form. For explanations are meant to provide *understanding*; how far they do so depends partly on their superficial linguistic features. In particular, *perspicuity* is an explanatory virtue, but one can make an explanation less perspicuous by superficial linguistic changes, such as randomly substituting synonyms for synonyms in some but not all occurrences or altering the order of exposition in confusing ways. Thus the helpfulness of explanations will be sensitive to such superficial linguistic features. But if our assessments of the truth-value of 'because' statements derive from how helpful we find the corresponding (putative) explanations, those assessments of truth-value will also be sensitive to superficial linguistic features. Such sensitivity can be built into the truth-conditions of 'because' statements only by giving them metalinguistic meanings, which for present purposes are irrelevant. Thus, as a guide to the truth-value of 'because' statements (read non-metalinguistically), the helpfulness of the corresponding explanations is at best an imperfectly reliable heuristic.

Similar considerations may help to clarify what is going on with (5A/B) and (6A/B). When we assess statements of the form 'S brought it about that Y', we may be envisaging a (putative) causal explanation of 'Y' in which S plays a central, active role. When extra conjuncts are added to 'Y', we expect the causal explanation to include a causal explanation of them too (otherwise, why bother to make them explicit?), still with S in a central, active role, on pain of inadequacy. For extra conjuncts such as 'John was self-identical'

and 'Edward was male', we realize that no such explanation can be included, and so are disappointed in the envisaged overall explanation, with negative effects on our assessment of the truth-value of the statement 'S brought it about that Y'.

Explanatory considerations may also be at work in Fine's original example, (1A/B). For we may expect something's essence to be fundamental to explanations of its other properties (perhaps combined with auxiliary hypotheses). We expect the name 'Socrates' to figure in good explanations of his non-essential properties. We do not expect the description 'a member of {Socrates}' to figure in such explanations, where it would normally introduce irrelevant complications. Thus, if we are using explanatory considerations as a heuristic for essentiality, that might well incline us to attribute opposite truth-values to (1A) and (1B). But explanatory considerations are sensitive to superficial linguistic features, and so are at best a fallible guide to the underlying metaphysics. Consequently, even if one premise of Fine's argument is false, it could easily strike us as true.

The hyperintensionalist 'revolution' is typically motivated by alleged counterexamples to intensionalism like (1A/B)–(6A/B), uncritically taken at face value. Such a basis is taken to justify adding multiple degrees of freedom in hyperintensionalist theorizing.

The most egregious case is 'impossible world semantics', where a world is any set of object-language sentences, however inconsistent or incomplete (though consistent and complete sets count as worlds too). The sentences true at such a world are all and only its members. In selecting a model, one must choose not only which possible worlds to include, but also which impossible worlds. By contrast with possible worlds semantics, determining which atomic formulas are true at a world in no way constrains which truth-functions of those formulas are true at the world: the modeller has complete freedom to decide. Indeed, this kind of impossible worlds semantics abandons the constraint, central to most formal semantics, of semantic compositionality, for the semantic profile of a complex sentence in a model is not determined by the semantic profiles of its constituent expressions in the model and the way they are put together. For instance, two formulas X1 and X2 may be true at the same worlds in a given model, even though their negations ¬X1 and ¬X2 are true at different worlds in the model. Similarly, the conjunctions X1 & Y

and X2 & Y may also be true at different worlds in the model. From a scientific modelling perspective, this looks paradigmatically like too many degrees of freedom.

Other forms of hyperintensional semantics are less profligate in adding degrees of freedom. For example, some hyperintensionalists (such as Fine himself) use forms of *truthmaker semantics* which respect the constraint of semantic compositionality. Nevertheless, they still multiply degrees of freedom.

First, a model associates each atomic sentence with a set of truthmakers, a subset of the overall set of states in the model. These truthmakers are not intended to do the modal work of possible worlds, so they add degrees of freedom not compensated by any subtraction of degrees of freedom associated with possible worlds, since the latter are in effect still needed.

Second, even when the admissible sets of truthmakers are subject to structural constraints (such as closure under a fusion operation associated with the model), in a typical model there are still many more admissible sets of truthmakers to choose between for a given atomic sentence than the two standard truth-values.

Third, since there seems to be no way to determine the truthmakers for ¬X in terms of the truthmakers for X, the semantics typically has a model assign each sentence *falsemakers* as well as truthmakers, so the truthmakers of ¬X can be the falsemakers of X and *vice versa*. That satisfies the letter of semantic compositionality, if not its spirit. However, since the set of falsemakers is independent of the set of truthmakers, this effectively doubles the number of degrees of freedom associated with the choice of a set of truthmakers for a given atomic sentence.

Thus even truthmaker semantics vastly increases a model's degrees of freedom. This is entirely typical of hyperintensional semantics.

In short, the self-proclaimed hyperintensional revolution involves multiplying degrees of freedom in order to explain data which may well be unreliable. That looks like a classic case of overfitting. Disturbingly, there seems to be no awareness within the hyperintensionalist camp that the programme carries these warning signs of bad science. The data are uncritically accepted, and the extra degrees of freedom are uncritically welcomed as increasing flexibility.

Consequently, the methodological challenges are not even being addressed. Of course, they *might* turn out to be just teething problems. If the programme can achieve enough explanatory success, it may eventually be vindicated. But, by normal scientific standards, merely accommodating data by multiplying degrees of freedom constitutes little in the way of explanatory success. It comes too cheap. Thus, on present evidence, the so-called hyperintensional revolution may well be a spectacular case of overfitting.

From this perspective, if analytic philosophy is to move up to the next level of methodological sophistication, it must take steps to avoid overfitting, by becoming less profligate in adding degrees of freedom, and more critical in assessing its data. For that purpose, talk of 'philosophical intuitions' is worthless. Instead, we should be asking ourselves what heuristics might have produced those data, and how reliable those heuristics are likely to be. That will be, not a revolution, but a reform in philosophical method. I hope that *Disputatio* will play a role in this reform.

Timothy Williamson
University of Oxford
New College, Oxford OX1 3BN, U.K.
timothy.williamson@philosophy.ox.ac.uk

References

Fine, Kit. (1994). "Essence and Modality". *Philosophical Perspectives 8*: 1–16.
Kripke, Saul A. (1979). "A puzzle about belief". In *Meaning and Use*, ed. by Avishai Margalit. Dordrecht: Reidel. pp. 239–83.
Williamson, Timothy. (2020). *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford: Oxford University Press.