Chapter Two of Timothy Williamson, *Overfitting and Heuristics in Philosophy* (OUP)
(draft of 23.4.2023)

Overfitting and Degrees of Freedom

1. *Error-fragility*

On the most naïve reading of Karl Popper's philosophy of science, scientific theories are falsifiable, but not verifiable. A scientific theory can be falsified, for it is a universal generalization, to which a particular negative instance, a counterexample, can be observed. But the theory cannot be verified, for however many particular positive instances are observed, they are all jointly consistent with a particular negative instance, which can be observed in the future, so they are all jointly consistent with the theory's negation.

Few contemporary philosophers of science accept that crude picture. By normal scientific standards, the theory of the circulation of blood *has* been verified. Of course, it has not been verified in the sense of having been *conclusively proved by the highest conceivable standard*, but then no scientific theory has ever been falsified either in the corresponding sense of having been *conclusively refuted by the highest conceivable standard*. After all, mistakes in observation are possible, and sometimes actual—through misperception and misinterpretation, incompetence and deceit, and so on. Scientific observation requires skill; things can go wrong. That is why scientists want important experiments to be repeated several times, preferably by different scientists in different laboratories. The reputation of the scientific team performing the experiment matters too—but not all reputations are deserved.

Imagine a scientific community proceeding as if naïve falsificationism were correct. As soon as someone reports an observation inconsistent with a scientific theory, the theory is trashed as refuted for all time, and the community never returns to it. If such a community ever entertains a correct theory, it is in serious danger of sooner or later throwing it out on the basis of a mistaken observation, and never returning to it. In the terminology of Joshua Alexander and Jonathan Weinberg (2014), such a methodology is *error-fragile*. A single error is liable to have catastrophic repercussions.

Even if the community raises its standard for accepting an observation report by demanding repeatability, that does not fully solve the problem of error-fragility. For when an experimental result is repeatable, something could still be wrong: the experimental design itself may be flawed; a crucial systematically interfering factor may have been overlooked, may even be unknown to the whole scientific community, so scientists are not measuring correctly what they think they are measuring. Deriving testable predictions from a scientific theory typically depends on auxiliary hypotheses too, for instance about how the experimental apparatus works, so a false prediction may result from falsity in an auxiliary hypothesis rather than falsity in the theory under test, as Pierre Duhem pointed out long ago.

Naïve falsificationism is inadequate in practice, not just in theory. It is a bad methodology, and it is not what scientists do. They rarely treat one observation report as refuting a theory. Even when no specific error has been identified, an apparent counterexample to an accepted theory may be treated as a mere *anomaly*, in something like Thomas Kuhn's sense, in the expectation or hope of resolving it eventually, unless an alternative theory nicely accommodates both the apparent counterexample and the other data (Kuhn 1962).

Of course, mathematics is one area of science where a single counterexample does indeed constitute a decisive refutation—once it has been proved by normal mathematical standards to be a counterexample. Even then, the validity of the proof can be contested, and time may be needed for the mathematical community to reach a consensus. In natural science, the methodological situation is often much messier.

Many contemporary philosophers follow a methodology uncomfortably close to naïve falsificationism. They use thought experiments rather than real-life experiments, but that does not solve the problem of error-fragility. A philosophical theory is put forward, in the form of a necessitated universal generalization. The theory—say, an account of knowledge as justified true belief—implies something about a specific possible case—say, that a Gettier case would be a case of knowledge. Pre-theoretically, we judge ('observe') that it would *not* be a case of knowledge. The case is then treated as a counterexample to the theory, which is therefore treated as refuted. In principle, treating our pre-theoretic capacity to judge what would obtain in thought experiments as a source of knowledge is not inherently problematic, just as treating our capacity to observe what obtains in real-life experiments as a source of knowledge is not inherently problematic.[1] But, equally, our pre-theoretic capacity to judge what would obtain in thought experiments is not *infallible*, just as our capacity to observe what obtains in real-life experiments is not infallible. In both natural science and philosophy, our fallibility in classifying examples suffices to make the naïve falsificationist methodology problematic, because it offers no proper means for identifying our errors and correcting them.

The standard methodology for employing thought experiments in philosophy is not maximally naïve, for it does not treat a lone philosopher's verdict on a thought experiment as refuting a philosophical theory. Idiosyncratic errors pose little threat to standard philosophical methodology, because the intellectual community will usually not adopt them—unless they are made by a charismatic or powerful figure in an intellectual sub-community with a very deferential culture. Normally, a verdict on a thought experiment will be generally accepted in philosophy only if most of those whose work it affects find it independently persuasive, without collusion. This means that thought experiments in philosophy are in effect required to be repeatable.

One could in principle worry about selection effects, where acceptance into a philosophical sub-community depends on sharing the received verdicts on some key thought experiments. That may indeed happen sometimes. But the overall trend in experimental philosophy over recent years has been to find that professional philosophers' verdicts on most thought experiments match the verdicts of lay people fairly well, once proper controls are in place by the standards of current experimental psychology—for example, to check subjects' comprehension of the hypothetical scenario (Mortensen and Nagel 2016, Knobe 2021). In crude terms, there is increasing evidence that the received verdicts in philosophy on thought experiments are mostly the natural human verdicts, irrespective of ethnicity and gender.

Nevertheless, the natural human verdict on a thought experiment is still a human judgment; it is not guaranteed to be *true*.

Chapter One explained a potent source of repeatable errors in verdicts on thought experiments: humanly universal heuristics. Such heuristics have limitations, and may even be implicitly inconsistent, so no interpretation makes all their deliverances true. The persistence heuristic lures us into false verdicts on sorites series. The suppositional heuristic generates false verdicts on conditionals. Disquotational heuristics do likewise in semantic paradoxes and Frege puzzles. If we take those verdicts at face value, following a naïve falsificationist methodology, we may as a result dismiss as refuted true theories of vagueness, conditionals, truth and falsity, and propositional attitudes. Merely scrutinizing the alleged counterexamples very hard will not solve the problem, especially if the scrutiny itself employs the very heuristic in question. The problem of erroneous data from thought experiments is not just potential; it is actual. Some natural human verdicts on thought experiments are false.

The problem is not confined to philosophy. Semantics as a branch of linguistics employs a similar methodology. Much of its evidence comes from natural human verdicts on sample sentences, envisaged as uttered in hypothetical circumstances. Those verdicts may be mediated by fallible heuristics, rather than issuing directly from semantic facts somehow directly available to the speaker. Linguists as well as philosophers are interested in the semantics of vagueness, conditionals, and propositional attitude ascriptions. They face many of the same methodological challenges.

The method of hypothetical cases, applied in a naïve falsificationist spirit, faces the problem of error-fragility in practice, as well as in theory, when verdicts on thought experiments result from humanly universal heuristics of limited reliability. The scientific analogy is not to a poorly executed experiment giving a single false negative, but rather to a repeatable experiment whose design ignores a hidden source of systematic error.

The scientific analogy does not support any proposal to abandon the method of hypothetical cases in philosophy. On the contrary, errors in data from thought experiments should no more motivate philosophers to give up using thought experiments than errors in data from real-life experiments motivate natural scientists to give up using real-life experiments. Instead, what the scientific analogy supports is the search for controls to mitigate the problem, by reducing error-fragility. We cannot realistically hope to prevent errors in the data from ever occurring. What we can realistically expect hope for are methods which will enable us both to prevent such errors from doing too much damage, and eventually to identify and correct the errors. In seeking such methods, we should at least consider how the problem of erroneous data is treated in natural science.

To be scrupulous, we should note a terminological subtlety. The analogy between erroneous data from real-life experiments and erroneous data from thought experiments presupposes that verdicts on the latter count as 'data' in the scientific sense. A verdict is like an attempted measurement of the truth-value of a proposition about the hypothetical scenario, somewhat as quantitative data are like attempted measurements of the value of some physical quantity. The term 'data' in this scientific sense is by no means precise. 'Data' are sometimes defined as 'facts', yet the occurrence of errors in data is acknowledged, even though there are no false facts.[2] As I will use the term 'data', some data are indeed false, so data are not facts. Hence data need not be evidence, given that nothing true is inconsistent with the evidence

(Williamson 2000: 200-202). For present purposes, we can think of the data as *prima facie* evidence.

Of course, thought experiments are by no means our only source of evidence in philosophy. On my view, anything we know is part of our total available evidence, in philosophy as elsewhere. But for present purposes, we can focus on thought experiments as the relevant source of evidence, in developing the analogy between data in natural science and data in philosophy. In any case, the best test of an analogy is to try it out.

2. *Data fitting*

In considering the error-fragility of a naïve falsificationist methodology, it is natural to focus on the case of a given data point as an apparent counterexample to a given theory. The question is whether one can rely on that data point. One may treat error as conceivable, though unlikely. But that narrow focus makes the problem look less urgent than it really is. For the overall situation is typically that we have a large body of data, which we are trying to design a theory to fit. Even if each single data point is probably correct, it is often almost certain that the whole data set contains at least one incorrect data point. In natural science, a given data set may include millions of data points. In philosophy, the numbers are obviously much smaller, but a branch such as epistemology has still accumulated scores of thought experiments, any of which may be used to test a given theory. It also has real-life examples, such as chicken-sexers, who can unreflectively but reliably classify chickens on sight as male or female. For simplicity, I will focus on thought experiments. In effect, philosophers are often trying to design a theory to fit such a body of data. Even if one is legitimately optimistic about the reliability of standard verdicts in philosophical thought experiments, it would be very rash to assume that *none* of the standard verdicts is incorrect, either through a heuristic's limitations or for some other reason. In philosophy as well as natural science, a reasonable assumption is that we are trying to design a theory to fit a set of data points not all of which are correct.

We cannot solve the problem simply by resolving to be more careful in coming to our verdicts on thought experiments. Careless errors in thought experiments are usually picked up quite quickly. Many standard thought experiments have been mulled over by the philosophical community for decades. A resolution to take more care would probably make little difference. In any case, if our verdicts are the products of fallible heuristics, taking more care might simply involve applying the heuristic more carefully, when the problem is with the heuristic itself. We cannot realistically expect to make ourselves into error-free thought experimentalists, any more than natural scientists can realistically expect to make themselves into error-free real-life experimentalists.

Instead, we need an error-robust methodology, which enables us to identify and correct our errors after we have made them, rather than vainly trying to ensure that we never make them in the first place. More than that, we need to *learn from our mistakes*, by understanding what went wrong and becoming less likely to make such mistakes in the future. We can make progress by considering how curve-fitting works in natural science.

To keep things simple, imagine that we are studying a physical variable y as a function of another physical variable x. The values of x and y in given units are real numbers. We measure the value of y for many different values of x, and graph the results. The aim is to define a mathematical equation (a curve) for y in terms of x which goes as close as possible to the points on the graph. As it turns out, that can be done *perfectly*. Although the number of points to fit may be large, it is still finite, say n. Then one can always find a polynomial equation of degree $n - 1$ that goes exactly through all the data points:[3]

$$y = a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n$$

Here the coefficients a_1, a_2, …, a_{n-1}, a_n are real numbers, parameters selected to fit the data. Since the equation is defined by these n independent parameters, the equation (or model) is said to have *n degrees of freedom*—n moving parts, as it were. By hypothesis, some of the data points are incorrect—something went wrong in the process of measurement. Even if there are no systematic errors in the data, there is still random noise. The curve goes through all the data points, irrespective of whether they are correct or incorrect.

What happens when new data points are obtained? Now we have a larger total number of data points, but it is still finite, say $n + k$. Almost certainly, the old curve does not go exactly through the new data points; in other words, the new data falsify the old theory's predictions. After all, we already know that the equation is incorrect, because it goes exactly through an incorrect data point, and so gives an incorrect value of y for that value of x. If the old curve went exactly through the new data points, that would be amazingly good luck, unless the errors are very systematic, since the new data points would have to fit in exactly with the old errors. But one can still find a polynomial equation of degree $n + k - 1$ that goes exactly through all the old data points and all the new ones. It will have $n + k$ independent parameters, b_1, b_2, …, b_{n+k-1}, b_{n+k}, so the new model will have $n + k$ degrees of freedom. Usually, the new polynomial will behave quite differently from the old one, especially for extreme values of x, where the old polynomial's behaviour will be dominated by that of its largest term, a_1x^{n-1}, while the new polynomial's behaviour will be dominated by that of *its* largest term, b_1x^{n+k-1}, which will be quite different. The new curve is also very likely to have more humps and dips than the old one. Although the new curve will coincide with the old one at the n original data points, the old curve will typically not be a good approximation to the new one elsewhere.

This process is repeated every time new data come in. The overall result will not be gradual convergence to the correct equation, since the degree of the polynomial always increases. Instead, there may be increasingly wild oscillation. All this involves large failures of prediction at each stage.

The pathology just described is not merely hypothetical. Something like it, in a milder form, is a familiar kind of bad science. Scientists call it 'overfitting' (Forster and Sober 1994). Textbooks of model selection warn against it (see for example Burnham and Anderson 2010). Overfitting is well known to result in unstable theorizing and predictive failures.

To some philosophers, the term 'overfitting' may sound like a contradiction in terms. Fitting the data is a good thing; how can one have too much of a good thing? Even granting that one's data set probably contains errors, one might still feel that since it is the best one has to go on, one can do no better than to fit one's theory to it as closely as one can. But bitter scientific experience shows how unlikely that approach is to end well.

For scientists, a key symptom of overfitting is an increase in degrees of freedom. A common platitude is 'With enough degrees of freedom, you can model anything'. To a philosopher, that may sound like welcome flexibility, but it is not intended in that spirit. Rather, the point is that if one has given oneself so much flexibility that one can model anything, then one can smoothly accommodate any errors in one's data, so no difficulty will occur to warn one of a potential error, and one will receive no warning that something is amiss. Not even the most grossly erroneous data point will stick out as anomalous. No data point will be a suspicious outlier, because one's curve will go through them all. A standard form of scientific criticism is that a model has too many degrees of freedom. If increasing the number of degrees of freedom is treated as cost-free, the likely result is unstable theorizing under the influx of new data. Too much flexibility, too much freedom, is a bad thing.

Scientific consensus strongly favours parsimony in degrees of freedom, even at the cost of a looser fit with the data. A large part of the rationale is that allowing oneself less flexibility to fit the data will tend to make incorrect data points show up as outliers, and so let underlying patterns emerge more clearly; the distorting effect of errors in the data is reduced. For example, when scientists use a polynomial, they like its degree to be as low as reasonably possible (linear is best), without totally flattening the data, to minimize the number of coefficients. This is a far more realistic and error-robust strategy than the hopeless aim of trying to be so careful that there are no incorrect data points in need of identification. It is a better strategy for achieving a reasonable level of predictive success.

As Malcolm Forster and Elliott Sober (1994) have argued, the problem of overfitting helps explain and justify scientists' preference for simple theories over complicated ones, for the number of degrees of freedom roughly measures the complexity of a model. By restricting themselves to comparatively simple theories, they make the data harder to fit, and so reduce the threat of overfitting.

Some philosophers have argued that simplicity is a virtue only in theories at the most fundamental level in physics and metaphysics (Sider 2016, to which Williamson 2016 replies). But that squares neither with scientific practice nor with its theoretical rationale. The threat of overfitting is just as serious in non-fundamental sciences such as geology, biology, and economics, and in non-fundamental branches of physics, as it is in fundamental physics, and biologists, economists, and non-fundamental physicists are just as keen to keep the number of degrees of freedom low. Even detectives prefer simple explanations of the evidence. One of the problems with conspiracy theories is that their complexity rapidly increases as more and more people with varying motives have to be notionally recruited into the conspiracy to explain how it managed to remain secret.

In the latter examples, measuring complexity by the number of degrees of freedom is admittedly a stretch. In truth, the standard definition of degrees of freedom in terms of independent parameters is less rigorous than it may sound. It is right in spirit, and often works in scientific practice, but it is not fully general or precise. After all, Cantor showed that there is a one-one correspondence between the real numbers and the *n*-tuples of real numbers for any given positive natural number *n*. Consequently, one can encode any ordered *n*-tuple of real numbers in a single real number, and thereby encode a model with *n* real-valued parameters (*n* degrees of freedom) in a model with just one real-valued parameter (one degree of freedom).

Even if one sticks to functions standardly used in natural science, one can fit all sorts of data with a wave-like sine function specified by just three parameters, for its amplitude, frequency, and phase, by making the frequency high and so the waves close enough together. The match may be perfect even though the data show no overt sign of wave-like behaviour and also perfectly match a simple cubic equation. Absent any background theoretical reason for expecting wave-like behaviour, preferring the sine function would seem scientifically bizarre. Yet cubic equations have four degrees of freedom, which is more than the sine function's three. Thus, the simple criterion of the number of independent parameters in the model is too crude to capture scientific practice exactly. Rather, it is a useful scientific *heuristic*, an imperfectly reliable sign of something subtler.

The vaguer but deeper lesson is that if we make fitting the data too easy, by helping ourselves to such a wide range of options that any supposed data will find a match, we also make ourselves easy victims of the data, because the process will not alert us to any defects or outliers in them. Flexibility has costs as well as benefits. We can still talk of 'too many degrees of freedom', understanding the phrase in that less formal way. Knowing how to recognize when there are too many may be a local matter of enculturation and experience in a given sub-discipline, depending on what is needed to achieve a reasonable level of predictive success in that area.

The informal understanding of 'degrees of freedom' also facilitates generalizations to philosophical methodology, since in philosophy we rarely deal with numerical equations, or fitting quantitative data, or data sets large enough for statistical significance. In most cases, we cannot expect literally to *count* the degrees of freedom in a philosophical theory, since deciding what to count as its 'independent parameters' would involve too many semi-arbitrary choices, though we can come close in more formal areas of philosophy.[4]

3. *Overfitting in philosophical analysis*

Comfort with a succession of increasingly complex theories is easily observed in the still-continuing twentieth-century tradition of providing would-be 'conceptual analyses', or just 'analyses', for philosophically significant terms of ordinary language such as 'know', 'cause', 'mean', and 'free', or for the concepts they are supposed to express. For example, in the case of 'know', one can see the complexity proliferate over a decade by leafing through the pages of Shope (1983). The adjective in the term 'analytic philosophy' has been closely associated with that tradition of analysis.

Since the analysans on the right-hand side of the analysis was supposed to be necessary and sufficient for the analysandum on the left-hand side, counterexamples could be to either the alleged necessity or the alleged sufficiency. Notoriously, alternating spirals grew of ever more complex analyses and ever more complex hypothetical counterexamples, each analysis provoking counterexamples and each counterexample inspiring revised analyses. If the counterexample was to the necessity of the analysans for the analysandum, showing the analysans to be too strong, one could weaken it by adding an extra disjunct. If the counterexample was instead to the sufficiency of the analysans, showing it to be too weak, one could strengthen it by adding an extra conjunct. But making the analysans weaker in one place often made it too weak somewhere else, and making it stronger somewhere often made it too strong somewhere else. Conjunctions of disjunctions and disjunctions of conjunctions started to emerge. The whole process was reminiscent of the tradition in Ptolemaic astronomy of adding epicycles whenever a new discrepancy with observation was found.

In retrospect, it is striking how little resistance there was for so long to the ramifying complications. From inside the tradition, it just felt like discovering more and more hidden complexity in ordinary concepts. The analogy with degenerating research programmes in natural science (in the sense of Lakatos 1970) may have been occluded from practitioners by their understanding of themselves as engaged in the *a priori* conceptual work of analysis, sharply contrasted with the *a posteriori* empirical work of science. From outside the tradition, it looks like a classic case of overfitting, with the typical symptom of adding ever more terms—here in the form of conjunctions or disjunctions—and the resulting theoretical instability and predictive failures in new cases.

A distinctive aspect of the case was that the postulated complexity was attributed not to the world at large but specifically to the cognitive resources of ordinary people, whose concepts or meanings were supposedly being analysed into more basic terms. As the conceptual structure became ever more complex, it became ever less plausible to regard it, even metaphorically, as if it were written into lexical entries for the target words in an ordinary person's head. Yet the complex structure had to be implicit in the ordinary use of the term, and somehow available to *a priori* reflection, even though normal speakers of the language, presented with the proposed analysis, would typically have great difficulty in so much as comprehending the analysans, and even more in guessing whether or not it corresponded to their use of the analysandum. Unease about the intended cognitive status of analyses found early articulation in the 'paradox of analysis' (Langford 1942): if the analysis is correct, the analysans expresses the same concept as the analysandum, so they differ only verbally, so the analysis is trivial; thus no analysis is non-trivially correct. The paradox continued to niggle, with no agreed solution, but also without doing much to slow the growth of the analysis industry.

If the complex structure of the analysans is supposed to play some cognitive role in the process of real-time thinking with the analysandum, questions of computational feasibility arise, which were never properly addressed. From an evolutionary perspective, it is hard to understand how the near-ubiquitous use of concepts such as 'know' and 'cause' (or 'make') in ordinary thought could fail to be counter-adaptive if they really had the apparently *ad hoc* complex structures the analyses attributed to them. Anyone who has tried working out whether such a philosophical analysans applies to a given hypothetical example will have experienced what a tricky and time-consuming task it is, comparable to a lawyer's job of applying a complicated piece of legislation to a given case. The easy fluency with which ordinary folk apply words like 'know' and 'make' in real time would be near-miraculous. On the other hand, if the complex structure of the analysans is *not* supposed to play some cognitive role in the process of real-time thinking with the analysandum, the intended status of the analysis becomes still more mysterious, given that it is supposed to be an analysis of a concept with which we think.

Why should philosophers even *expect* philosophically interesting concepts to have analyses? Analysis is not supposed to be infinite; it is supposed to bottom out somewhere—why not straight away, at least for philosophically interesting concepts, and perhaps for most or all concepts (Fodor 1998)? In early analytic philosophy, the programme of analysis was motivated by much more general assumptions. For example, Bertrand Russell proposed the Principle of Acquaintance: '*Every proposition which we can understand must be composed wholly of constituents with which we are acquainted*'; he called it 'the fundamental epistemological principle in the analysis of propositions containing descriptions' (Russell 1910-11, his italics). Given Russell's extreme empiricist conception of acquaintance, on which we are not acquainted with ordinary material objects, the analysis of almost any proposition is forced to go far below the surface to reach a level of constituents with all of which we can be acquainted. But later analytic philosophers pursued programmes of analysis whose prospects of success were not supported by any such wider vision.

In recent decades, the ideology of 'concepts' and 'conceptual analysis' has come under increasing pressure. Our firmest grip on concepts comes from the words supposed to express them, but then we need an answer to the question: when does a word W used in a context *c* express the same concept as a word W* used in a context *c**? No really helpful answer is available. Identity in reference is presumably insufficient, since most theorists of concepts agree that 'water' and 'H_2O' can refer to the same stuff without expressing the same concept. One may be told that W in *c* expresses the same concept as W* in *c** if and only if the condition for understanding W in *c* is the same as the condition for understanding W* in *c**, but that is unhelpful because the conditions for linguistic understanding are so loose and vague. Alternatively, one may be told that W in *c* expresses the same concept as W* in *c** if and only if W in *c* is governed by the same rules as W* in *c**, but that will turn out to be circular because the rules themselves are individuated in terms of their constituent concepts. And so on. Without a workable theory of identity conditions for the concepts words are supposed to express, we lose methodological control of inquiry into the concept expressed by a given word W in a given context *c*, since we cannot tell which uses of W are irrelevant because they express a different concept. For purposes of this book, we need go no deeper into the disarray of concept theory: not even a theory of concepts in good working order would make the methodological issues of overfitting and heuristics go away.[5]

As enthusiasm for conceptual analysis has waned, and metaphysics has revived in analytic philosophy, many of the philosophers who still seek analyses understand their project as a metaphysical rather than conceptual quest. They seek to analyse causation itself, rather than the concept of causation, or freedom itself, rather than the concept of freedom. Correspondingly, the standard of success for an analysis is just for the analysans to be *metaphysically* necessary and sufficient for the analysandum, rather than *conceptually* necessary and sufficient, and perhaps also for the analysans to be *metaphysically* prior to the analysandum in some sense, rather than *conceptually* prior. The focus has switched from *how* we are thinking to *what* we are thinking about.

In practice, the change has been less radical than it sounds. Objections to a given conceptual analysis were usually cases where the analysans held without the analysandum, or *vice versa*; since they usually look metaphysically as well as conceptually possible, they also serve as objections to the corresponding metaphysical analysis. Similarly, objections to a given metaphysical analysis are usually cases where one holds without the other; since they usually look conceptually as well as metaphysically possible, they would also serve as objections to the corresponding conceptual analysis. Again, objections to conceptual priority can often be recycled as objections to metaphysical priority, and objections to metaphysical priority can often be recycled as objections to conceptual priority. For such reasons, switching the operative standard from conceptual analysis to metaphysical analysis does little to improve the track record of analysis. The series of analyses do not look convergent; rather, they exhibit the kind of theoretical instability and predictive failures associated with overfitting.

Even if we grant for the sake of argument that everything is somehow metaphysically reducible to an absolutely fundamental level, it does not follow that the reduction of the target phenomena—causation, freedom, knowledge, meaning, whatever—to the fundamental level will be mediated by an initial reduction of them to other phenomena at the highly non-fundamental level characteristic of a philosophical analysans—for instance, for typical analyses of knowledge, at the level of belief, truth, justification, causation, counterfactuals, and the like. The theoretical instability and predictive failures in the programme's track record is evidence that there is no such mediation.

At this point, a friend of philosophical analysis might try to recruit the considerations about heuristics and overfitting to its aid, by arguing that some original, simple analyses may have been right all along, with the apparent counterexamples being mere artefacts of fallible heuristics.

In the case of the original justified true belief analysis of knowledge (JTB), for example, the suggestion would be that knowledge really is simply justified true belief, while the standard negative verdicts in Gettier cases reflect a limitation of a universal human heuristic for ascribing knowledge (Weatherson 2003 makes a related suggestion). The idea should not be dismissed out of hand. However, to be made fully convincing, rather than remaining just another application of a generic sceptical argument, such a defence of JTB would need, first, to specify what the guilty heuristic is, second, to provide independent evidence that we really use such a heuristic, third, to show how the proposed heuristic delivers a negative verdict on Gettier cases, and fourth, to explain why a charitable interpretation of our practice of using 'know' and similar words nevertheless picks out as its referent justified true belief, thereby falsifying our verdicts on Gettier cases, rather than a relation more directly related to the heuristic and absent from Gettier cases, thereby verifying our verdicts. I am not aware of any promising attempt to meet any of those challenges.

The case is worth dwelling on, to see why JTB is *not* an example of a natural, elegant, explanatory hypothesis prematurely dismissed as a result of a glitch in a heuristic. Instead, JTB already shows signs of overfitting; it is an early precedent for the post-Gettier tradition of *ad hoc* analysis-building.

A defence of the original JTB analysis must employ the original understanding of 'justified', on which justified *false* belief is possible, as Gettier (1963) emphasizes. After all, if only true beliefs can be justified, the truth conjunct in the analysans is redundant, which no defender of JTB intends. Similarly, to understand justification in terms of knowledge would be contrary to the spirit of JTB. Thus, although justified true belief may be necessary and sufficient for knowledge on a more demanding normative understanding of justification, that does not amount to a defence of JTB (for such a more demanding normative conception of justification see Williamson 202W).

A different way to assess the plausibility of JTB is by noting that knowledge is a central focus for our ordinary thought and talk about cognitive matters (Williamson 2000, Nagel 2014): is justified true belief a good candidate to play that role? To take one case, our best ordinary understanding of the actions of non-human animals and young children often involves attributing knowledge to them—for example, they need to know where other agents and other relevant things are—but normative questions of justification in the sense of JTB seem out of place and digressive as applied to non-human animals and young children, who are not responsible agents (see also Kornblith 2002). The distinction between knowing and not knowing is much more primitive than the distinction between having and lacking justification for a belief. To see what happens when one starts from the latter distinction rather than the former, we can look at the work of numerous contemporary epistemologists of an internalist bent, who do indeed regard the distinction between having and lacking justification for a belief as the right starting-point for epistemology. The category of knowledge is typically marginalized in their work, as is the category of justified true belief, which from their perspective looks like an odd ill-hybrid of quite disparate factors.

Internalist approaches to epistemology also generate their own distinctive objections to JTB. For example, they typically insist that the external reliability of perception can vary independently of internal phenomenology and justification, and allow that perceptual knowledge requires some level of perceptual reliability, as well as internal justification. On such assumptions, an internalist can easily construct a pair of good and bad cases, identical internally in one's phenomenology, justification, and belief, and externally in the truth-value of a proposition *p* about the environment, although one's perception is reliable only in the good case, so that one knows *p* in the good case but not in the bad case. In such a set-up, JTB automatically fails, for since one knows *p* in the good case, by JTB one has a justified belief in *p* in the good case. Hence, by the internal identity of the two cases, in the bad case too one has a justified belief in *p*, and by hypothesis *p* is still true, so one has a justified true belief in *p*, but one does not know *p*. Thus, JTB fails in the bad case. In other words, the very epistemological outlook that vindicates JTB's order of analysis, by treating justification and belief as more fundamental than knowledge, undermines JTB for other reasons. Consequently, JTB is an ill-motivated theory.

We can also test JTB in a more abstract structural way, by seeing how it plays out in simple models within the formal framework of doxastic logic. The upshot is that although JTB is not a disjunctive analysis in the usual sense, for its analysans is a conjunction, not a disjunction, it has a subtler disjunctive effect, for the propositions known according to JTB in a model can be shown to be exactly the disjunctions of a true proposition and a proposition believed with justification, one disjunct to fix the T conjunct and another to fix the JB conjuncts (Williamson 2013, 2015). Although some of the assumptions built into the models may be unrealistic, such as the closure of justified belief under disjunction (one has a justified belief in a disjunction whenever one has a justified belief in a disjunct of it), an analysis that behaves so awkwardly under simplifying assumptions is not likely to behave much better when more complications are permitted. Thus, JTB makes knowledge into a rather artificial, gerrymandered category, not at all a natural candidate for reference, and so unlikely to be meant by the word ‘know’, since its fit with the use of the word is also poor, as Gettier cases show.

The point of invoking formal models here is not to make thought experiments redundant, but rather to make the rejection of JTB more robust, by basing it on a consilience of different methods, each pointing to the same conclusion. We have multiple reasons for regarding JTB as a bad theory.

The objections to JTB do not mechanically generalize to other proposed philosophical analyses, but they are suggestive. For example, the formal modelling illustrates how the apparently unifying effect of conjunctive analyses can be more apparent than real, when the conjuncts are not fruitfully related to each other. A conjunction of miscellaneous factors is also a clue to overfitting, since it suggests too many degrees of freedom.

The method of checking theories by calculating how they play out in simplified formal models is capable of being applied far more widely than it currently is in philosophy. It often provides valuable structural information, not least because it displays a theory’s consequences over a whole space of propositions, not just one proposition at a time. Most basically, it is a test of the theory’s consistency, both in itself and with elementary background constraints. Where such tests are possible, they should be applied. They can save much time and energy wasted on hopeless theories that fail the test.

The programme of philosophical analysis is now most associated with the long-running quest for necessary and sufficient conditions in more basic terms for philosophically central yet non-logical words like ‘know’, ‘mean’, ‘cause’ and ‘free’. Overfitting became rife, for philosophers were not taught to minimize degrees of freedom. But, I conjecture, it has not led to the rejection of correct analyses on the basis of misjudged examples, because in these cases there are no correct analyses of the kind sought.

4. *Overfitting in semantics*

The formal semantics of natural languages is pursued in both departments of linguistics and departments of philosophy, with similar methodologies. It makes an interesting test case for issues about overfitting and degrees of freedom, for several reasons.

First, the formal framework lends itself to counting degrees of freedom. Typically, the semantics is implemented in formal models. Typically, a model assigns semantic values to expressions of the natural language under study relative to various parameters. For simplicity, we can treat the function mapping each sequence of values of those parameters to the semantic value of the expression relative to that sequence as the *meaning* of that expression in the model. The theorist chooses a set of parameters so as to enable the semantics to be *compositional*, in the sense that, in any model, the meaning of a complex expression is determined as a function of the meanings of its simpler constituent expressions and how they are put together. That helps explain how users of the language can understand newly encountered sentences composed of previously encountered words.

As a first approximation, we can equate the number of parameters in the formal framework with the number of degrees of freedoms of a model. Really, the situation is more complicated, because a model must assign a meaning separately to each atomic expression of the language—roughly speaking, to each word—so each atomic expression adds another degree of freedom to the model. However, since rival formal frameworks tend to agree that each atomic expression must be assigned its own meaning, and on which expressions are atomic, we can assume that all these degrees of freedom cancel each other out when formal frameworks are compared, leaving only differences between the parameter sets themselves, that is, between the sets of all meanings available in a given framework.[6]

The formal semantics of natural languages lends itself to discussion of overfitting in another respect too: it is strongly data-driven. A typical driver for theory change is that someone identifies examples in some natural language which the old framework seems unable to handle (a background methodological assumption is operative: that the formal framework should be *universal*: suitable for all human natural languages). The examples are typically in the form of sample sentences, often with native speaker judgments as to whether they could be correctly uttered in given hypothetical circumstances. In effect, the data are verdicts on elementary thought experiments. One or two data points of that kind may be taken to motivate a revision of the formal framework.

The formal semantics of natural languages also provides a good test case for accounts of overfitting because, historically, some highly successful revisions of a formal semantic framework *have* indeed taken the form of adding a new parameter, increasing the number of degrees of freedom. For instance, Saul Kripke revolutionized the semantics of modal logic in the period 1959-63 by enhancing models with a new parameter for a 'possible world', in what is sometimes described as the start of the 'intensional revolution'. This development is worth describing in some detail.

Before Kripke's innovation, there were *extensional* models for non-modal predicate logic. Each non-modal model provides a set of individuals to be the domain of quantification, an extension over the domain for each atomic predicate of the formal object-language, and a referent in the domain for each individual constant. The truth-value of each formula of the object-language in the model is then defined compositionally relative to each assignment of values to all variables, in the usual way. For simplicity, I will henceforth leave this relativity to assignments tacit; it only makes a difference to the final truth-value for formulas with free variables. One can define a conclusion to be a *logical consequence* of a set of premises if and only if the conclusion is true in every model in which every premise is true. Similarly, a formula is a *logical truth* if and only if it is true in every model. Famously, there are formal proof-systems for first-order non-modal logic which can be proved *sound* and *complete*, in the sense that if a conclusion is provable in the system from some premises then it is a logical consequence of the premises (soundness), and conversely, if the conclusion is a logical consequence of the premises then it is provable in the system from those premises (completeness).

The non-modal object-language can be expanded to a modal object-language by the addition of modal sentence operators such as ◊ (informally read as 'possibly) and □ (informally read as 'necessarily'). Modal operators do not fit into the standard extensional framework because the extension of a formula in such a model is simply its truth-value, and modal operators are not truth-functional: the truth-value of the output is not a function of the truth-value of the input. For example, if a formula α is false in a model, $\Diamond\alpha$ may be either true or false in the model, and if α is true, $\Box\alpha$ may be either true or false. An increasingly urgent question in the 1940s and 1950s was how to adapt models for non-modal logic to modal logic, preferably so as to enable analogous soundness and completeness theorems to be established for appropriate formal proof-systems for first-order modal logic.

A simple, natural, and economical strategy is to treat the modal operators as generalizing over the extensional models themselves. More specifically, for any formula α, $\Diamond\alpha$ is true in a model if and only if α is true in some model, and $\Box\alpha$ is true in a model if and only if α is true in every model. Thus, possibility is understood as truth in some model, and necessity as truth in every model. If we regard extensional models as demystified possible worlds, then possibility is equated with truth in some possible world, and necessity with truth in every possible world. Rudolf Carnap (1947) pursued a strategy along these lines, using syntactic 'state-descriptions' rather than extensional models, but to very similar effect. He even compared his state-descriptions to Leibniz's possible worlds (which were ideas in the mind of God). The consequent reduction of modality to syntax was attractive from the perspective of Carnap's logical positivism.

However, for various technical reasons, the Carnapian approach worked poorly (Williamson 2013: 75-80). In particular, no formal proof-system is sound and complete for the logical consequence relation it generates. Although many logicians in the 1950s contributed to the search for an alternative approach to the model theory of modal logic, it was Kripke who took the decisive step, in effect by sharply separating the role of possible worlds from that of models. He defined a new kind of model. Such 'Kripke models' are *intensional* models. A Kripke model is equipped with a set W, required only to be nonempty, whose members can informally be thought of as possible worlds, although that plays no role in the development of the model theory proper, which involves only mathematical and syntactic reasoning. In a model, the members of W are in effect the available values of the world parameter. The model assigns each 'world' (each member of W) a set of individuals to be its domain of quantification. The model also assigns each atomic predicate an *intension*, a function mapping each 'world' to the predicate's extension at that world, defined over the appropriate domain. Formulas are assigned truth-values compositionally relative to each 'world' (they are also relative as before to an assignment of values to variables, which we can ignore for present purposes). For any formula α, $\Diamond\alpha$ is true at a 'world' w if and only if α is true at some 'world' in W, so possibility is understood as truth in some world, and $\Box\alpha$ is true at w if and only if α is true at every 'world' in W, so necessity is understood as truth in every world (in the simplest version of the semantics). The model also singles out one member of W, which is informally understood as the actual world. A formula is evaluated as true in the model, without relativization to a 'world', if and only if it is true at the 'actual world' of the model, but the 'non-actual worlds' are still needed to determine whether modal formulas are true at the 'actual world' of the model, since the modal operators are interpreted as quantifiers over worlds.

The simple structures just described are quite restrictive, because they all validate the strong modal logic S5, on which nothing is contingently possible or contingently necessary. The formulas $\Diamond p \rightarrow \Box\Diamond p$, $\Diamond\Diamond p \rightarrow \Diamond p$, and $p \rightarrow \Diamond p$ are all theorems of S5, but fail on many interpretations of the modal operators. For example, when $\Diamond$ is interpreted in terms of easy possibility (and $\Box$ as $\neg\Diamond\neg$), $\Diamond\Diamond p \rightarrow \Diamond p$ is invalid, because even when one can get from A to B by an easy step, and one can get from B to C by an easy step, it does not follow that one can get from A to C by an easy step. When ◊ is interpreted in terms of permissibility, or compatibility with what one believes, or the past or future, $p \rightarrow \Diamond p$ is invalid, for when something happens, it does not follow that its happening is permissible, or compatible with what one believes, or past, or future. Kripke therefore equipped his models with an *accessibility* relation R between 'worlds'. Formally, R can be any binary relation over W. The generality over 'worlds' is then restricted by accessibility, in the sense that $\Diamond\alpha$ is true at w if and only if α is true at some world to which w has R (possibility is truth in some accessible world) and $\Box\alpha$ is true at w if and only if α is true at every world to which w has R (necessity is truth in every accessible world). To invalidate $\Diamond\Diamond p \rightarrow \Diamond p$, one allows R to be non-transitive; to invalidate $p \rightarrow \Diamond p$, one allows R to be non-reflexive. The accessibility relation makes the formal framework much more flexible—and in doing so adds another degree of freedom.

Kripke's work had a profound influence on philosophy. The apparatus of possible worlds soon became a standard part of an analytic philosopher's toolkit, a convenient framework for use in thinking and talking about all sorts of topics. Philosophical theses were increasingly formalized in modal rather than non-modal terms. Quine's arguments had put modal language under a cloud of suspicion, with dark threats of incoherence, especially when the possibility or impossibility at issue concerned individuals themselves, irrespective of how they were designated, and so resisted paraphrase in terms of the consistency or inconsistency of sentences. Although Kripke's formal semantics by itself gives little specific information on *which* such *de re* modal claims are true, it does call the bluff of those threats of incoherence, and demonstrates very clearly that there is no purely logical obstacle to the meaningfulness of *de re* modal claims. Natural versions of his semantics also validate some significant structural principles with a metaphysical edge, such as the non-contingency of identity and distinctness. The result was in effect to give the green light to substantive theorizing about modal metaphysics, in which Kripke himself played a leading role.

On the technical side, the perspicuous formal structure of Kripke models lends itself to mathematical investigation, and the model theory of modal logic became a flourishing branch of mathematical logic. It also found numerous applications in other disciplines, often using models equipped with whole families of accessibility relations, each associated with its own modal operators. For example, in computer science, modal logic is applied to the study of indeterministic computing, where the members of W are interpreted as the possible states of the computer, and each programme is associated with an accessibility relation which one state has to another if and only if running the programme *can* take the computer from the former state to the latter (Pratt 1976 is a seminal paper on *dynamic logic*, Troquard and Balbiani 2022 a recent survey). Kripke models are also standard in epistemic and doxastic logic, where the knowledge and belief operators are indexed to agents, and one state has a given agent's accessibility relation to another if and only if, when the agent is in the former world, for all they know (or believe) they are in the latter world. Such models are used in theoretical economics and computer science, as well as in formal epistemology (Fagin, Halpern, Moses, and Vardi 1995).

Here, our interest is in the contribution of Kripke models to the formal semantics of natural languages. Although the object-language for his model theory was a formal language, its operators ◊ and □ were generally understood to be formal representations of modal operators such as 'possibly' and 'necessarily' in natural languages. Discussions of modal metaphysics often moved seamlessly in and out of languages for quantified modal logic and natural languages. From a linguistic perspective, the most common modal terms are auxiliaries such as ‘must’, ‘can’, ‘could’, ‘may’, ‘might’, ‘would’, ‘should’, ‘ought’, and so on. They are commonly used to make highly contingent claims, for which Kripke models are far more natural than a Carnapian framework: the accessibility relation can be as local as desired. Linguists soon applied a world parameter to the semantics of modal auxiliaries in natural languages (Kratzer 1977 was especially influential). More generally, the use of an intensional framework with a world parameter for the formal semantics of natural languages became standard.

In short, introducing a world parameter to semantic models proved immensely fruitful in logic, philosophy, linguistics, and beyond. It was clearly a progressive move. Adding a degree of freedom is not *always* bad.

The treatment of context-dependence offers another case of the fruitful addition of new parameters to a formal semantic framework. The occurrence of terms whose reference varies with context is not in doubt: obvious examples include demonstratives like this’, ‘that’, ‘then’, ‘there’, and ‘they’, and other indexicals like ‘I’ and ‘now’. Such context-dependence is not ambiguity: that the word ‘I’ refers to me when uttered by me and to you when uttered by you is explained by the same linguistic rule; we use the word with the same linguistic meaning. Although it is controversial how far such context-dependence extends beyond the obvious cases, the need for a proper semantic treatment of it is clear. Such a treatment will require contextual parameters, in order to formulate general linguistic rules such as the rule of reference for ‘I’.

The seminal recent account of the semantics of context-dependence is by David Kaplan (1989). One might hope that the parameters needed to handle intensional operators could also be used to track context-dependence: for example, that Kripke’s world parameter for modal operators and an analogous time parameter for temporal operators would in effect track shifts in the world and time of the context. But Kaplan showed that it is not so. Take the word ‘tomorrow’. As uttered on a given day D, it refers to the day after D, D+1. Imagine someone saying on D ‘When it’s tomorrow, I’ll feel better’. To handle the phrase ‘when it’s tomorrow’, which applies ‘when’ to the sentence ‘it’s tomorrow’, the compositional semantics must evaluate ‘it’s tomorrow’ with respect to different times, to determine which of them it is true at. But that is quite different from determining when one can truly say ‘it’s tomorrow’, for the answer is: never (trick cases aside). To get the right result, the semantics must evaluate ‘it’s tomorrow’ *as uttered on day D* with respect to other days: as uttered on day D, ‘it’s tomorrow’ is true with respect to day D+1. In Kaplan’s terminology, one must distinguish the *context of utterance* (on day D) from the *circumstance of evaluation* (on day D+1). The reference of ‘tomorrow’ is fixed in the context of utterance, and then carried over to the circumstance of evaluation. Since the circumstance of evaluation varies independently of the context of utterance, separate time parameters are required for each, and likewise for world parameters. That is how ‘tomorrow’ manages to be a *rigid designator*, even though its designation varies over time: if the context of utterance is held fixed, its designation remains the same while the circumstance of evaluation is varied. By contrast, context-dependence is variation in reference (or designation) when the circumstance of evaluation is held fixed while the context of utterance is varied; that is why ‘tomorrow’ is context-dependent. Kaplan uses the distinction between context of utterance and circumstance of evaluation to implement his general theory of content and character, where the content of an expression in a given context is what it contributes to the propositions expressed by sentences containing it as uttered in that context, while its character is the function mapping each context to the expression's content in that context. The content of 'I' in a given context is normally the speaker of that context. The character of 'I' is what remains constant across contexts, a good candidate for its linguistic meaning.

In short, a proper semantic treatment of context-dependence would hardly be possible without something like the distinction between context of utterance and circumstance of evaluation, and the consequent multiplication of parameters.

Both Kripke's work and Kaplan's were, and still are, paradigms of successful innovation in formal semantics. That may well have given semanticists the impression that this is just what progress in semantics looks like: introducing one or more new parameters into the formal semantic framework to explain linguistic data that could not be explained otherwise. That is just what one would expect from a Kuhnian perspective on semantics: scientists recognize solutions to new problems by their similarity to paradigmatic solutions of old problems. The trouble in this case is that if one keeps introducing new parameters into the semantic framework, thereby increasing degrees of freedom, one will sooner or later sink into overfitting.

We cannot reasonably expect a universal formula for when to stop adding parameters, much though some philosophers might demand one. As so often in science, it requires experience and good judgment. But one must at least recognize the problem, and not regard the introduction of a new parameter as cost-free. Each new parameter makes overfitting more likely.

A semanticist might object that if a new parameter is needed to explain the data, it would be a dereliction of duty *not* to introduce one. But that is a generic reply, which can always be offered in defence of overfitting. We must ask whether the data really are all correct, and whether the new parameter really is needed to explain them. If introducing a new parameter is regarded as a paradigmatic form of progress in semantics, or at least as methodologically low-cost, then there is little incentive to probe the data for errors, or to keep seeking an alternative explanation for the data within the current framework.

For example, relativism as a view in contemporary semantics involves adding a parameter to the circumstance of evaluation for something like a *standard of assessment*, in order to explain data about predicates of personal taste and other phenomena (Lasersohn 2005, MacFarlane 2014). That the linguistic phenomena have been correctly described, and can be explained only by adding an assessment parameter to the circumstance of evaluation, is by no means obvious (Cappelen and Hawthorne 2009). For instance, I say to Ana 'Rhubarb is disgusting'; Ana says to me 'Rhubarb is delicious'. In a sense, we disagree; in a sense, we are both right. But if I spoke in a context where the relevant reference class comprised just me, while she spoke in a context where the relevant reference class was just her, then we both spoke truly, and the apparent disagreement was merely verbal, as if I had said 'I like rhubarb' and she had replied 'I don't'. By contrast, if we both spoke in a joint context where the relevant reference class comprised both me and her, then the disagreement was real, but we both spoke tendentiously and falsely. Once such confusions as to the operative context have been cleared up, the invocation of a new parameter in the circumstance of evaluation may have nothing left to explain.

As in the case of philosophical analyses, adding too many parameters is not the only form of over-complication in semantics. One can also overfit by overusing the standard contextual parameters, for example, to gerrymander a complicated contextualist semantics on which content varies with context in *ad hoc* ways, or by diagnosing context-dependence more widely than necessary.

Not all semanticists accept that simplicity is a theoretical virtue in the semantics of natural languages. Perhaps the concern is that meaning in natural languages may just be very complicated. But that too is just an instance of a generic form of scepticism about simplicity as a theoretical virtue. A scientist in any branch of science, accused of overfitting, can respond that what they are investigating may just be very complicated. Indeed, in semantics, apart from the usual need to avoid overfitting, there is the additional concern that an over-complicated semantic framework may impose infeasible computational burdens on ordinary speakers. In contemporary semantics, one often sees very complex semantic accounts presented with no apparent sense that their complexity might be a theoretical cost.

A current test case is the research programme of *dynamic semantics* (not to be confused with Pratt's dynamic logic). In slogan form, the central idea is that 'meaning is context change potential'. Dynamic semantics is motivated by phenomena such as cross-sentential anaphor. For example, I can say 'Samuel kicked a stone' and later add 'It rolled into a ditch'. Together, my two statements are equivalent to 'Samuel kicked a stone, which rolled into a ditch', but it is broken into two. In a standard formalization of my original statement, the phrase 'a stone' would introduce an existential quantifier, with no implication of uniqueness; he may have kicked several stones. My use of the pronoun 'it' in my second statement is anaphoric on 'a stone', so one wants to formalize it with a variable bound by the existential quantifier, but that does not work because the quantifier's scope is confined to my original statement. Although this is not a straightforward counterexample to a non-dynamic framework, it is unclear how to handle it within such a framework. By contrast, dynamic semantics in effect extends the scope of 'a stone' over the whole subsequent discourse. Dynamic semantics is a generalization of standard truth-conditional semantics, in the sense that the latter can be recovered as a special case of the former, but dynamic semantics is significantly more complex and flexible than standard truth-conditional semantics. As another example, 'A and B' is not in general equivalent to 'B and A' in dynamic semantics, since the second conjunct is processed with respect to a context updated on the first. A recent introductory survey of dynamic semantic emphasizes its flexibility as a framework (Nouwen, Brasoveanu, van Eijck, and Visser 2022), but, as we have seen, the obverse of flexibility is overfitting. The jury is still out on whether any linguistic phenomena are explicable *only* by dynamic semantics. The risk is that dynamic semantics turns out to be another manifestation of overfitting.

A more specific case is the semantics of conditionals. In their attempts to do justice to the complex ways we use conditionals in natural language, semanticists have offered a wide variety of complex semantic and pragmatic accounts of those conditionals. Elsewhere, I have argued in detail that humans' primary heuristic for assessing such conditionals is the suppositional heuristic described in Chapter One, and that it is implicitly inconsistent

(Williamson 2020). *No* semantics will validate all aspects of our use of 'if'. Consequently, the search for a semantics that *does* validate all those aspects is condemned to overfitting. Instead, we do better to accept that our use of 'if' is flawed, the data cannot all be taken at face value, and the semantics of 'if' must be related to our use of it less directly. That opens the way to rehabilitating the simplest of all semantics for 'if', the material, truth-functional interpretation. How far that approach can be generalized to other problem cases for the semantics of natural languages, I leave to the reader as an open question.

5. *Overfitting in logic*

On a popular stereotype of logic, it is not in the business of fitting data, and so cannot be guilty of overfitting. Instead, logic is imagined as laying down ground-rules without which our language could not function, let alone express our hypotheses and the data we test them on. But such preconceptions about logic find no support in the actual practice of disputes between proponents of rival logics in the same natural language.

Most famously, Hilary Putnam once argued that data from 'two-slit' experiments in quantum mechanics can best be explained on the hypothesis of a failure in the distributive principle of classical logic, that *P* and (*Q* or *R*) entails that (*P* and *Q*) or (*P* and *R*) (Putnam 1969). Putnam later withdrew his conjecture, and the programme of 'quantum logic' is generally regarded as a failure, at least in its attempt to explain the puzzling data (Putnam 2012). However, even if Putnam's reasons for making the proposal were confused, there was never a transcendental proof that any such proposal *must* be confused—unless one counts an argument that relies on the classical logical principles in question. Imagine that tomorrow a team of leading experts in logic and quantum mechanics announces that they have found a better way to explain the data from experiments in quantum mechanics on the hypothesis of a failure in some generally accepted principle of classical logic. A philosopher who tells them that they must be confused, because their proposal violates the rules of the language, would not have much credibility. One might be sceptical of their proposal on inductive grounds, because so many similar proposals have turned out badly in the past, but such scepticism is itself data-driven.

Whether a given allegedly logical principle has the status of a rule of some natural language is a question about that language, to be settled on the basis of evidence by the normal standards of linguistics. Such evidence might include data on what speakers of the language are or are not willing to say in various speech situations. The evidence would need to discriminate between linguistic rules and regular theoretical principles to which speakers are deeply committed. Strikingly, philosophers who ascribe the status of a linguistic rule to a logical principle tend to provide little or no linguistic evidence to support their claims.

In any case, ascriptions of exceptional linguistic status to logical principles are of scant dialectical use in defending a generalization against alternative logicians who deny that it *is* a logical principle. After all, if the generalization is *not* a logical principle, all kinds of data may be used against it. If someone asserts that Newton's laws of motion are laws of logic, a physicist may appropriately respond by providing experimental evidence against Newton's laws. If an alleged law of logic is false, it is not a law of logic. Thus, when critics of classical logic bring all kinds of data against it, the response that logic is not in the business of fitting data would be question-begging. Instead, if friends of classical logic decide to take the critique seriously, they have to get their hands dirty by engaging with the alleged counter-evidence and showing what in particular is wrong with it.

Very schematically, if the criteria for theory comparison are divided into simplicity, strength (informativeness), and fit with data, then proponents of alternative non-classical logics typically have no choice but to make their case on grounds of fit with data. For their alternatives are clearly neither simpler nor stronger than classical logic.

In practice, all kinds of phenomena have been wheeled out against classical logic. For instance, the law of excluded middle has been alleged to fail for the open future, the forgotten past, potential infinity, vagueness, semantic paradoxes, quantum physics, and so on. In each case, the critics willingly give examples where they take excluded middle to have unacceptable consequences. If they often leave it unclear exactly what evidence they are relying on in those cases, it is not for want of trying. But if friends of classical logic manage to show that it can accommodate such challenging phenomena, they have thereby enhanced the case for classical logic. Beyond that, both sides aim at more than mere accommodation: they want to treat the phenomena at issue in their preferred logic smoothly and elegantly, without resort to *ad hoc* devices.

More positively, proposed laws of logic gain support by identifying a common structural pattern in a mass of examples with diverse subject matters, unifying them by bringing them all under one illuminating generalization. For instance, the gradual identification of *modus ponens* as a logical principle in ancient Greece was a very significant intellectual achievement; as a general principle, it was not obvious all along (Bobzien 2002). Again, for many readings of modal operators, which principles of modal logic are sound remains unclear—especially for principles where modal operators occur embedded in the scope of further modal operators. Settling such questions is at least in part a matter of data-fitting (see also Ripley 2016). In a non-causal sense of 'explanation', we aim at an inference to the best explanation of the data.

Since logic is involved in data-fitting, it faces the issue of overfitting. Indeed, the issue takes some of the same forms for logic as we saw it take for semantics, for logical consequence is standardly defined by a generalization over semantic models. Standardly, a conclusion is defined to be a logical consequence of some premises if and only if every model of the premises is a model of the conclusion, in other words, the conclusion is true in every model in which every premise is true. Some variations on that theme are played for some non-classical logics, but they all involve generalizations over semantic models. Thus, changing the class of models by adding new parameters can change the logical consequence relation, at least for those connectives whose semantics is sensitive to the change.

For instance, Kripke's semantics for non-modal intuitionistic logic involves adding a new parameter with an associated reflexive, transitive accessibility relation. Informally, the picture is that the parameter's values are states of information, and one state is accessible from another when the latter extends the former. The semantics is tweaked so that every formula behaves monotonically: if a state of information verifies a formula, so does every extension of that state. Monotonicity requires a tweak to the semantic clause for negation ($\neg$): instead of the classical clause that a state verifies $\neg\alpha$ if and only if it does not verify α, Kripke's semantics requires a state to verify ¬α if and only if *no extension of* that state verifies α, in order to ensure that whenever a state verifies ¬α, so does every extension of that state. As a result, the semantics invalidates the principle of double negation elimination: in some cases, a state verifies ¬¬α without verifying α. For example, in a model with just two states, *s* and *s*+, where *s*+ extends *s*, and only *s*+ verifies an atomic sentence *p*, neither state verifies ¬*p*, so both states verify ¬¬*p*, so *s* verifies ¬¬*p* but not *p*.

Without adding new parameters, one can also increase flexibility by extending the range of values available for an old parameter, by adding or subdividing values. Standard semantic models are *bivalent*: at a given point in the model, each sentence is either true or false, and not both. In three-valued logic, the available values are typically truth, falsity, and 'neither'. In four-valued logic, they may be 'just true', 'just false', 'neither true nor false', and 'both true and false' (*sic*). Such flexibility is used to deal with semantic paradoxes such as the Liar. In response to paradoxes of vagueness, the values may spread out into a continuum, represented by the real numbers from 0 (perfect falsity) to 1 (perfect truth), allowing the truth-value of the vague sentence 'It's dark' to rise continuously at dusk and fall continuously at dawn. A multi-dimensional space of truth-values may be invoked to track the multi-dimensionally vague sentence 'It's a religion' in truth-value. As truth-values proliferate, so do the different possibilities for defining logical consequence in terms of them. In some three-valued logics, preservation of non-falsity determines a different consequence relation from preservation of truth. In some other many-valued logics, a conclusion is a logical consequence of some premises if and only if no model makes the conclusion worse in truth-value than every premise, and so on.

Conversational virtues may also be built into the definition of logical consequence, by adding new parameters to models. In relevance logic, validity requires the conclusion to be in some sense relevant to the premises, and models are complicated accordingly (see Anderson and Belnap 1975 and Mares 2004; see Burgess 1981 for a critique of the idea of 'fallacies of relevance'). Perhaps someone will propose *politeness logic*, whose models have a politeness parameter taking values in a totally ordered set, informally understood as a scale from the rudest to the politest. A model assigns each atomic sentence a rudeness-value. The rudeness-value of a complex sentence is the maximum (worst) of the rudeness-values of its atomic constituents. An argument is valid only if it is politeness-preserving, in the sense that no model makes the conclusion ruder than every premise (other conditions may also be necessary for validity). In politeness logic, the rule of disjunction introduction is invalid, for if a model makes the atomic sentence *r* ruder than the atomic sentence *p*, then it also makes the disjunction *p* ∨ *r* ruder than *p*, so *p* ∨ *r* is not a logical consequence of *p*.

The semantics for a logic can be complicated in other ways too. Most obviously, the semantic clauses for logical connectives can be gerrymandered to invalidate disliked principles. The possibilities are endless.

Revisions of classical logic are often presented as making for more flexibility. Usually, classical logic can still be recaptured from the proposed alternative as a special case. For example, restricting Kripke models for intuitionistic logic to those with only one state in effect collapses it back to classical logic. Restricting models for many-valued logic to those which assign each atomic sentence one of the two 'classical' truth-values has a similar effect. Thus, the non-classical model theory can be interpreted as recognizing all the possibilities recognized by classical model theory, and more besides. The non-classical logic is strictly *weaker* than classical logic, since some arguments in the object-language validated by all classical models are invalidated by some models in the non-classical semantics, whereas every argument validated by all models in the non-classical semantics is also validated by all classical models.

Another way of laundering weakness in a logic as a virtue is by saying that a weaker logic *makes more distinctions* than a stronger logic. Specifically, given two formulas α and β, $\alpha \leftrightarrow \beta$ may be a theorem of a logic L but not of a weaker logic L^-, though $\alpha \leftrightarrow \alpha$ and $\beta \leftrightarrow \beta$ are theorems of both. Then friends of L^- may say that it distinguishes between α and β, whereas L does not. Of course, the distinctness of the *sentences* α and β is not in dispute. The question is whether $\alpha \leftrightarrow \beta$ is *true* on all relevant interpretations. If it is, what is the virtue in being unable to prove it? Analogously, imagine a theory of arithmetic so weak that it lacks '2 + 2 = 4' as a theorem. The distinctness of the *terms* '2 + 2' and '4' is not in dispute. The question is whether 2 + 2 and 4 are the same *number*. If they are, what is the virtue in being unable to prove it?

Proponents of non-classical logic face the challenge of explaining the success of classical logic as the standard implicit background logic for proofs in mathematics for two and a half thousand years, by far the most severe test of any logic in human history. A popular strategy is to claim that the language of pure mathematics satisfies the special conditions for the recapture of classical logic from the preferred non-classical alternative. For example, those who reject excluded middle for languages with vague or meta-semantic vocabulary (such as 'true' and 'false') often accept it for the language of pure mathematics, which they take to lack such vocabulary. However, *applications* of mathematics in the natural and social sciences do involve vague or meta-semantic vocabulary, so the classical-recapture strategy arguably fails to explain the success of classical mathematics in scientific applications (for details see Williamson 2018). Such alternative logicians cannot escape as easily as they imagine from the daunting challenge of reconstructing mathematics for scientific applications from the starting-point of their weak non-classical logic.

On one view of logic, it is needed to play the role of a neutral arbiter of more substantive disputes in science or metaphysics. That view favours weak logics, because they are neutral on more questions. Strength in a logic tends to compromise its neutrality. But what counts as 'substantive' is never made clear. In any case, the view is hopeless because *any* proposed principle of logic can be attacked on scientific or metaphysical grounds, however mistaken they may be, and so is not neutral on those scientific or metaphysical issues. We have already seen examples of that, and they can be multiplied. Under the influence of Hegel, a metaphysician may claim that all change involves a contradiction. Even the anodyne structural principle that logical consequence is reflexive, so α is always a logical consequence of α, may be denied by a follower of Heraclitus, on the grounds that one can never grasp the same proposition twice. If a correct logic comprises only principles incapable of being challenged on scientific or metaphysical grounds, then the correct logic is empty.[7]

For present purposes, a key feature of weak logics is how unhelpful they are when we need to identify bad data. For a crude case, take dialetheist logic, which permits true contradictions. When a witness contradicts himself, the dialetheist is not best placed to see the problem. We might also be suspicious of a witness who states 'Not everyone present was invited' but refuses to accept 'Someone present was not invited', a consistent combination of attitudes in intuitionistic logic. Of course, alternative logicians may offer more roundabout reasons for doubting such testimony. More crudely, they may just say that, since classical logic is incorrect, we should not be relying on it in our attempts to identify bad data.[8]

Such responses on behalf of alternative logics push the question further back. Were concerns about overfitting given enough weight, or indeed any, when the case for revising classical logic was made? Increases in flexibility make data, including bad data, easier to accommodate, and so incur a significant methodological cost. Sensitivity to this cost is hard to detect in arguments against classical logic. Instead, one finds remarkable levels of implicit trust in unclear data—for example, when the failure of classical logic for vague languages is simply taken for granted. The problem becomes even more acute when the data can be explained as products of imperfectly reliable heuristics, as Chapter One explained they often can.

One reason why the analogy between alternative logics and prototypical cases of overfitting may have been missed is that there are also striking differences. Most notably, in curve-fitting, the old and new curves represent equally specific hypotheses. By contrast, alternative logics are usually less specific—weaker—than classical logic. They withdraw from some consequences of the old hypothesis, without adding new ones to compensate. That is quite different from what happens in the natural sciences. The analogue would be at best a kind of curve-fitting where one specifies upper and lower curves, the hypothesis being that the correct values lie in the band between the two curves. The analogue of weakening the logic would be pushing the two curves further apart, widening the band to include data-points outside the old band, thereby weakening the hypothesis. For some cases, the analogue would be even worse: just crossing out the old curve without proposing a replacement. With such a methodology, the result of erroneous data-points is not a more erroneous hypothesis but just a less informative one. Consequently, one will not get the kind of erratic instability and falsified predictions characteristic of classic overfitting, but just a loss of informativeness.

At a more general theoretical level, weakening a logic is analogous to revising a theory in physics by abandoning some of its general principles without replacing them by any alternatives of similar generality. Such a move in physics would look defeatist rather than progressive. Not only would it result in a less informative theory; it would fail to stress-test the crucial data by not trying to explain them on an alternative equally strong theory. That would have in common with overfitting an insufficiently critical attitude towards the data. Proposals for weakening classical logic implicitly treat the relevant data in a similarly uncritical way.

6. *Overfitting in philosophical model-building*

When scientists speak of degrees of freedom in a model, the models they have in mind are rarely semantic models such as those in the previous two sections, which assign semantic values to expressions of a language. Although semantic models are a special case of models in the general scientific sense, the scientists are unlikely to have semantic models in mind. Most models of natural or social phenomena imply nothing specific about the semantic values of linguistic expressions—though general methodological morals about models in science are indeed applicable and relevant to semantic models. In the predominant scientific sense, a model of a phenomenon is an intermediary object, which is intended in relevant ways to be easier to study than the target phenomenon itself, but structurally similar enough to it for insights about the model to reveal something about the target phenomenon (for a general treatment of modelling in science see Weisberg 2013). For the model to have well-defined parameters, it must be formally specified, typically in mathematical terms, for example by differential equations whose coefficients are parameters of the model. Informally, the equations may be conceived as characterizing the development of a closed system over time. By solving the equations analytically, or by approximating their effect for given initial conditions on a computer, scientists can often work out how the model develops, identify patterns, and tentatively transfer them to the target phenomenon. Models of pandemics and of global warming are of this general kind.

Almost always, the model is much simpler than the target phenomenon. For example, it may model a planet as a point mass. Without such massive simplifications, the model would be mathematically and computationally intractable, and so not fit for purpose. Unless the target phenomenon is the whole universe, just modelling it as a closed system—ignoring the possibility of interference from outside the system—is already a massive simplification: in practice, there is always some outside interference. Restricting degrees of freedom in models to avoid overfitting is another source of deliberate simplification. In a deterministic model, the implicit assumption that the parameters' values at one time jointly determine their values at any later time is also a simplification. For example, a biological model of predator-prey interaction may treat the sizes of the two populations at any time as jointly determining their sizes at any later time, ignoring obviously relevant factors such as age profile, interactions with a third species, the changing state of the environment, and so on.

Although models in natural and social science tend to be diachronic, that is inessential to the model-building methodology. Probability spaces in the mathematical sense are used to model various kinds of uncertainty, but are synchronic: they specify one probability distribution, not an evolving sequence of distributions. Many models of language in linguistics are synchronic. So is an electoral model of the relation between parties' share of the votes and their share of the seats in an assembly. One can learn about synchronic dependencies between the variables by varying the values of the variables, thereby comparing different models with the same overall structure.

Much progress in the natural and social sciences consists in building better models of natural and social phenomena. The new model may capture all the features of the target phenomenon captured by the old model, and more besides, without becoming mathematically or computationally intractable. That is quite different from the older paradigm of scientific progress as the discovery of new scientific laws. Most macroscopic systems and many microscopic ones are too messy and complicated to satisfy any distinctive universal generalizations, let alone laws, formulated in terms of such systems.

Philosophers sometimes try to hold onto a law-based conception of science and absorb such complications by invoking a category of '*ceteris paribus* laws', but it cannot do the required work. *Ceteris paribus*—other things being equal—planets are *not* point-masses, and the number of animals in a population does *not* vary continuously over time as a differential equation requires—it is discrete. We can determine by rigorous mathematics or computer simulation what holds in a model; by contrast, the assumption that a law holds *ceteris paribus* is far too vague for us to determine its consequences in any such way. Instead, the appropriate way to study those messy and complicated phenomena is often by building formal models rather than by seeking exceptionless or *ceteris paribus* laws. One may be able to conclude with some rough but robust generalizations drawn from the model, formulated with '*ceteris paribus*' qualifications (Weisberg 2013: 158-9, 167-8), but at the heart of the rigorous scientific action is the model, not a *ceteris paribus* law. The principles that define the model which explains the *ceteris paribus* law do not themselves hold *ceteris paribus*.

Of course, model-building in the natural and social sciences is informed by data. The curves in curve-fitting are more or less simple models. A complication at this point is that the target phenomenon is hardly ever a particular one-off event token, such as the Big Bang or global warming on Earth. Typically, the target phenomenon is a general *type* of event or process, such as the working of some kind of cell or bodily organ, or some kind of interaction between two species. Scientists aim to produce a model of the type, that is, a model generically of a token of the type, without there being any particular token of which it is a model—just as a diagram can be of a human heart without there being any particular human heart of which it is a diagram. Typically, the model will be based on data from many different tokens of the relevant type. Even if there are no *errors* in the data, there may still be *outlying* data, for instance from an abnormal heart. Such outlying data can distort the model, making it a worse model of a normal heart (similar issues would arise in semantics too if it were done in a similarly model-building spirit too). Precautions against overfitting can help avoid such distortions too.

Not all model-building concerns quantitative issues. For example, biologists want to understand the predominance of sexual reproduction, since asexual reproduction is also possible, and actually occurs in some cases. For such theoretical purposes, biologists use a model-building methodology. In particular, they build models to explore the hypothesis that sexual reproduction makes a population better able to adapt to changes in the environment. Such a model may schematically represent competition between a sexually reproducing population and an asexually reproducing population, with an initial distribution of genotypes, under intense selection from a rapidly changing environment. The hypothesis to be tested is that sexual reproduction is more conducive than asexual reproduction to variance in genotype, and so makes for more adaptability to environmental change (Weisberg 2013: 115-117, Crow 1992). The dynamics of such a model are encoded in simple mathematical rules, whose consequences over time can be computed. The rules are not intended to be a realistic description of actual events, but just to capture general qualitative features of competition between the two forms of reproduction. The rules are best kept as simple as possible, not merely for ease of computation, but because any unnecessary complication or *ad hoc* feature would risk rigging the model in favour of a desired conclusion. The aim is not data-fitting or quantitative prediction but general theoretical explanation. Still, the model is ultimately constrained by data: for instance, the dynamical rules for how the sexually reproducing population evolves must not run counter to what is known about the genetics of sexual reproduction.

In brief, simple formal models are used in the natural and social sciences to gain qualitative understanding of very general phenomena, especially phenomena whose instances are too messy and complicated to be governed by informative exceptionless laws at the relevant level. That is evidence that the model-building methodology would be of value to philosophy too. An example was already noted in section 3: the use of simple formal models to test the behaviour of the JTB analysis of knowledge. We should expect the model-building methodology to be much more widely applicable in philosophy than that. After all, the human world is fantastically messy and complicated ('Out of the crooked timber of humanity no straight thing was ever made'), and much of human philosophy concerns the human world: human minds and bodies, human knowledge and action, human thought and language, human art and science, human morality and politics, human identity through change and counterfactual variation. Many of those philosophical issues generalize to the wider world of non-human animals and perhaps even of artificial intelligence, but that wider world is fantastically messy and complicated too. Thus, philosophy spends much of its time and energy engaging with phenomena of just the kind better suited to model-building than to the quest for exceptionless laws.

The model-building methodology is already widely used in some areas of philosophy (Williamson 2017a). It predominates in formal epistemology, including Bayesian probabilistic models, models of epistemic and doxastic logic, and more, for both individual and social epistemology. Formal models from decision theory, game theory, deontic logic, voting theory, and evolutionary theory are sometimes used in moral and political philosophy. In metaphysics, the mereology of gunk—the theory of parts and wholes where everything has a proper part—is hard to think about accurately without a mathematical model. When semanticists of natural language state a formal semantic theory, it is usually for a toy model language.

Despite these examples, only a small proportion of contemporary philosophy uses a model-building methodology. One possible explanation is that only a small proportion of contemporary philosophy studies topics for which a model-building methodology would be useful. However, in my experience, there is also widespread ignorance and incomprehension of the model-building methodology amongst philosophers. Many take the falsity of its simplifying assumptions as sufficient reason to reject a model. Most graduate students in philosophy are neither encouraged to build or use models nor trained in how to do so. Consequently, the methodology has not so much been tried and found wanting as not tried. Almost certainly, it could usefully be applied more widely in philosophy than it has been so far, as some examples below suggest.

Unfamiliarity with the model-building methodology helps explain the marginalization of epistemic and doxastic logic in twentieth-century mainstream epistemology, despite the pioneering work of Jaakko Hintikka (1962). For decades, new developments in epistemic and doxastic logic came more from computer scientists and theoretical economists than from philosophers. Notoriously, standard models of epistemic logic validate a strong form of *logical omniscience* for knowledge: automatically, if one knows the premises of a deductively valid argument, then one knows the conclusion too (with no qualifications about knowing the entailment or having competently carried out the deduction). Similarly, standard models of doxastic logic validate the correspondingly strong form of logical omniscience for belief: automatically, if one believes the premises of a deductively valid argument, then one believes the conclusion too (again, with no qualifications). The obvious computational and reflective limitations of actual humans seem to provide innumerable counterexamples to logical omniscience for both knowledge and belief, which were taken to discredit epistemic and doxastic logic: whatever epistemic and doxastic logicians were studying, it was not what interested epistemologists, they thought. Significant opportunities for epistemology were lost. Logical omniscience could have been treated as a legitimate simplification for modelling purposes, in order to apply the formal framework of epistemic and doxastic logic to the rigorous exploration of other structural issues in epistemology. But such an outlook on model-building seems not to have been available in epistemology at the time. The required formal skills were also in short supply.

Later work in epistemic and doxastic logic showed how to avoid logical omniscience, and model agents' limited rationality, for example by allowing possible worlds epistemic or doxastic access to 'impossible worlds' where any set of sentences whatsoever can be the set of truths (Rantala 1982). Such models carry a cost, because they drastically increase degrees of freedom. The effect is that the model-builder has to put the agent's knowledge and beliefs into the model 'by hand', so stipulation largely replaces exploration, and many of the potential gains from model-building are lost. Instead of learning from the model, one is just taking out of it what one had explicitly put in.

Some functionalists in the philosophy of mind have argued that, on a proper understand of knowledge and belief, the failures of logical omniscience are illusory (Stalnaker 1984, 1999). The following chapters will discuss issues about logical omniscience and the individuation of the objects of knowledge and belief more deeply. In any case, technical work on blocking logical omniscience was scarcely noticed in mainstream epistemology, and did not in practice constitute a bridge between mainstream epistemology and epistemic and doxastic logic. Encouragingly, more recent years have seen more interaction between formal epistemology and mainstream epistemology, and consequently more use of the model-building methodology in epistemology by philosophers.

When the model-building methodology is applied in philosophy, issues of overfitting and degrees of freedom arise. We need to be on the alert for complication and *ad hoc* features as warning-signs of error or distortion. If an example convinces philosophers that a phenomenon can occur, but modelling it requires something reminiscent of gerrymandering, then the example may have been misinterpreted. Even if a phenomenon is genuinely possible, it may be such an outlier that complicating the model to allow for it makes the model less useful for many other purposes.

Often, the dialectic is more intricate. Here is a case from my own experience. I have been interested in using epistemic models to illustrate extreme failures of the 'KK' or 'positive introspection' principle that if one knows that P, one knows that one knows that P, and of watered-down versions of that principle—for instance, that if one knows that P, one is *in a position to* know that one knows that P. In models of epistemic logic, positive introspection provably corresponds to the transitivity of the accessibility relation for knowledge. I have used an example where one is looking from a distance at an unmarked clockface with a single hand, wondering what time it shows. By looking, one learns *something*, but not *everything*, about where the hand is pointing. In the simplest epistemic models of that situation, one can easily show, positive introspection fails drastically (Williamson 2014b). However, friends of positive introspection can approximate those models by slightly more complicated models where positive introspection *holds*, by arbitrarily selecting a coarse-grained partition of epistemic possibilities to determine the accessibility relation in the model. Similar tricks can be played with less grossly simplified models where positive introspection fails.

However, one can show that any model of the situation which validates positive introspection does so at the cost of a sort of *symmetry-breaking* (Williamson 2021d). The basic set-up, including all potential positions of the hand, has a symmetry induced by the underlying rotational symmetry of the clockface about its centre. But one can show that if an epistemic model respects that symmetry, in the sense that the underlying epistemic structure remains invariant under 'rotations' of the model, and the model meets basic epistemic constraints such as avoiding both scepticism and omniscience about where the hand is pointing, then the model violates positive introspection. Symmetry-breaking is not just inelegant; it is a symptom of an *ad hoc* intrusion. In effect, it means that the model postulates differences in associated epistemological structure between one point on the circumference of the clockface and another, even though the basic set-up does not require any such differences. Of course, in real life, the situation is doubtless not perfectly symmetric: our visual system may well embody a slight bias on one side or another. But for an epistemological theory to insist in advance that there *must* be such a bias or asymmetry in a situation whose basic structure does not impose one looks like gerrymandering. In particular, specifying which way the bias goes would require an extra degree of freedom. Thus, the methodology of model-building tells against symmetry-breaking, and so against positive introspection. In brief, to save positive introspection in this case, you must overfit.

None of this means that *all* epistemic models must invalidate positive introspection. On the contrary: there is often good reason to use epistemic models that validate both positive introspection and the much less plausible principle of *negative introspection*: if one does not know that P, one knows that one does not know that P, which corresponds to the accessibility relation being *Euclidean*, in the sense that all worlds accessible from a given world are accessible from each other. Negative introspection fails in very ordinary cases of confident error, where it is false that P, but one thinks that one knows that P: as a result, one does not know that P, but one also does not know that one does not know that P (in Donald Rumsfeld's phrase, an unknown unknown). Despite such clear counterexamples, it often makes good methodological sense to build negative introspection as well as positive introspection into a model. For many epistemic models are designed for exploring interpersonal effects, such as those which depend on the presence or absence of common knowledge (where everyone knows that P, everyone knows that everyone knows that P, everyone knows that everyone knows that everyone knows that P, and so on). Each agent in the envisaged situation has their own knowledge operator with its own epistemic accessibility relation. The best way to isolate the *inter*personal obstacles to common knowledge is by minimizing the *intra*personal obstacles to knowledge about one's own knowledge states, to avoid interference. That requires assuming that each agent satisfies both positive and negative introspection separately. Unfortunately, epistemic logicians have tended to treat positive and negative introspection not just as convenient simplifying modelling assumptions but as all-purpose epistemological dogmas, brushing aside all the epistemological objections to them. When the focus shifts to intrapersonal epistemology, background assumptions of interpersonal epistemology should be called into question. Which simplifying assumptions are legitimate depends on what questions one is using the model to think about.

7. *Summing up*

As we have seen, the methodological issues around overfitting and degrees of freedom are themselves messy and complicated, both in general and in particular in their application to philosophy. But that is no good reason to ignore them. Although it can be hard to know whether one is overfitting, we need to be alert to the danger. To that end, we should put somewhat more weight on simplicity and strength as theoretical virtues in philosophy, as elsewhere, and be somewhat less trusting than heretofore of data that tempt us to complicate and weaken our theories. Obviously, that does not mean that we should rush to the opposite extreme: having neglected simplicity and strength, we should not switch to neglecting fit with data. We need to be keenly aware that there is a balance to be struck, and that we have not been striking it right. Nothing about the nature of philosophy exempts us from the methodological exigencies that other theorists are used to working under.

Clearly, once we have a plausible explanation of potential errors in those data the theory fails to fit, the lack of fit becomes less serious. As we have seen, in philosophy and elsewhere, our reliance on a fallible heuristic may be central to that explanation.

In abductive methodology, theoretical virtues are often roughly summarized as simplicity, strength, and fit with data. Giving too much weight to any one or two of them at the expense of the remainder distorts our view of comparisons between theories, and is liable to end in error or triviality. By focusing on the more specific theoretical vice of overfitting, we can clarify our sense of what would be involved in properly implementing an abductive methodology in philosophy.[9]

Notes

[1] Chapter 6 of Williamson 2007 explains the legitimate role of thought experiments in philosophy as verifying counterfactual conditionals for use as premises in philosophical arguments. Williamson 2021a extends the defence of that account. Chapter 14 of Williamson 2020 streamlines and strengthens the account by replacing the Lewis-Stalnaker semantics for counterfactual conditionals with a simpler semantics on which they express contextually restricted strict conditionals.

[2] For quantitative data, 'correct' should be qualified by 'within the specified margin for error'.

[3] For simplicity, I assume that the value of *y* has been measured at most once for any given exact value of *x*. I also assume that the number of parameters required to fit a polynomial to noisy, partially incorrect data will be the same as the number of data points. These assumptions are typically correct.

[4] Recent successes of AI at tasks such as face recognition and text prediction have been taken to mandate a re-evaluation of accepted wisdom about overfitting, since they have been achieved by using almost unlimited numbers of degrees of freedom. The programmes are typically trained on vast sets of typically accurate data, and their success is measured by predictive accuracy rather than theoretical understanding. One cannot plausibly argue from such cases that natural scientists have been misguided in their strategy of limiting degrees of freedom. Since philosophical inquiry is much more similar in its aims and methods to scientific inquiry than to face recognition or text prediction, it would be similarly implausible to argue from the successes of AI that philosophers should not be guided by a strategy of limiting degrees of freedom. By training an AI programme on a vast set of data on human answers to questions about a wide range of thought-experimental scenarios, with no restriction on the number of degrees of freedom, one might indeed manage to get the programme to answer questions on further thought-experimental scenarios in a human-like way. But that would not advance philosophical understanding of the issues the scenarios were intended to probe: the AI is just doing what humans can already do, perhaps in a slightly different way. It is not producing an intelligible theory of the philosophical subject matter. Worse, when we have evidence that the data is infected by heuristic-induced errors, it does not tell us which natural human verdicts on the scenarios are *true*, and which *false*. In this respect, the recent successes of AI are not of much help to philosophy.

[5] The status of conceptual analysis is of course closely related to that of the analytic-synthetic distinction, famously attacked by Quine 1951. My critique of analyticity or conceptual truth is developed in Williamson 2007, extended with many replies to objections in Williamson 2021a; it differs from Quine's strategy by not depending on scepticism about semantics. For a nice contrast between extreme optimism and extreme pessimism about the prospects for conceptual analysis, see Jackson 1998 and Fodor 1998.

[6] Some semantic theories posit complex logical forms for syntactically simple expressions, such as ‘cause to die’ for ‘kill’, but the choice of logical form for a given such expression itself constitutes at least one degree of freedom. A further complication is that a formal framework may treat some atomic expressions (such as ‘is’, ‘not’, ‘or’, ‘and’, ‘if’, ‘some’, and ‘all’) as ‘logical constants’ with a fixed interpretation, so models need not assign them meanings separately. Formal frameworks can differ from each other in which atomic expressions they treat as logical constants. However, most atomic expressions of a natural language are not plausibly treated as logical. In practice, the whole situation is often much messier than indicated in the text—which is typical of model-building. I have aimed to provide a reasonable first approximation.

[7] For a more detailed critique of the idea of logic as a neutral arbiter see Williamson 2014. For more on strength as an abductive virtue in logic see Williamson 2017. For how confusion between logic and metalogic has led to scepticism about strength as an abductive virtue in logic see Williamson 202ZZ. Much of the literature on disagreement in logic is vitiated by similar confusions, for example between disagreement in logic on whether $\forall P\ (P \vee \neg P)$ (the law of excluded middle) and disagreement in metalogic on whether ‘$\forall P\ (P \vee \neg P)$’ is a logical truth.

[8] Naïve empiricist attitudes to data may lead to errors in logic as they seem to have done elsewhere in philosophical theorizing. According to Ernst Cassirer, in Renaissance Italy ‘empiricism leads not to the refutation but to codification of magic’ and to ‘empirical magic’, through its respect for miscellaneous empirical reports of magic (1963: 151-2), whereas rationalists swept such reports aside in their quest for a uniform mathematical theory of nature. In present terms, the empiricists were guilty of overfitting.

[9] See also Williamson 2016c on abductive philosophy.