TIMOTHY WILLIAMSON
University of Oxford
timothy.williamson@philosophy.ox.ac.uk

MODEL-BUILDING AS A PHILOSOPHICAL METHOD*

abstract

The method of building simplified formal models of phenomena under study is widespread in contemporary natural and social science; much scientific progress consists in the provision of better models. A model-building methodology has also been used with success in analytic philosophy, for example by Carnap in his development of intensional semantics. Arguably, philosophers have overlooked how much progress their discipline has made through their failure to conceive it in model-building terms. By using the method more extensively, they can overcome the fragility to error inherent in the naïve falsificationist methodology on which many analytic philosophers rely.

keywords

models, idealization, progress, Carnap, intensional semantics, falsificationism, Popper, error-fragility

* The ideas in this paper were presented at the 2018 conference on *The Methods of Philosophy* at Cesano Maderno; I thank participants for valuable discussion. Elsewhere I put this account of model-building in philosophy in the context of a general account of philosophical method (Williamson, 2018) and develop it further (Williamson, 2017).

Phenomenology and Mind, n. 15 - 2018, pp. 16-22
DOI: 10.13128/Phe_Mi-24968
Web: www.fupress.net/index.php/pam

ISSN 2280-7853 (print) - ISSN 2239-4028 (on line)

1. Models in natural science

Many natural scientists aim at a distinctive kind of progress which philosophers are just starting to recognize as an appropriate aim for them too.

The stereotype of scientific progress is discovering a new law of nature. Such laws are meant to be universal generalizations about the natural world, holding without exception for all times and places, by some sort of necessity: nice, if you can find one. However, most natural science studies messy complex systems – cells, animals, planets, galaxies – which are hard to characterize by universal laws. What laws must hold of all tigers, for example? 'All tigers are striped' won't do, because there are albino tigers. 'All tigers are four-legged' won't do either, because there are three-legged tigers, and so on. 'All tigers are animals' is true, but doesn't get us far. Although tigers obey the fundamental laws of physics, like everything else in nature, that won't console a biologist who wants to say something specific about living things as contrasted with elementary particles and stars. If we keep watering down our initial attempts, we may eventually reach something exceptionless, but the danger is that it will be too weak and uninformative to be of much interest. This isn't just a problem about animals. Complex systems of all shapes and sizes tend to be messy and unruly.

To manage the problem, scientists have revised their objectives. Instead of seeking universal laws about complex systems, they build simplified *models* of them. Occasionally these are physical models: water running through a sand tray to model a river eroding its banks, a construction of colored rods and balls to model a DNA molecule. More typically, the models are abstract, defined by mathematical equations which describe how a hypothetical system changes over time. The hypothetical system is vastly simpler than the real-life systems of interest, but still has a few of their key features. The strategy is to analyse the behavior of the hypothetical system mathematically, in the hope that it will simulate some puzzling aspects of the real-life systems' behavior, and thereby cast light on them (Weisberg (2013) provides a good introduction to the philosophy of scientific model-building).

For example, you might wonder why a population of predators – say, foxes – and a population of prey – say, rabbits – keep oscillating, though rises and falls in one do not coincide with rises and falls in the other. A key point is that, holding other things equal, the more foxes there are, the more rabbits get eaten, but the more rabbits there are, the more fox cubs survive. One can write down differential equations that express the rate of increase or decrease of each population in terms of the current number of predators and prey. They are known as the Lotka-Volterra model. In most ways, it is grossly over-simplified: it ignores changes in the vegetation the rabbits feed on, changes in the tendency of humans to hunt foxes and

rabbits, variations among foxes, variations among rabbits, and so on. Since such factors make a difference, the equations are not universal laws. Indeed, they couldn't be, since for mathematical reasons the change in population is treated as continuous, even though in real-life it changes in whole numbers: when one of 200 rabbits dies, the number goes straight down to 199, with no intermediate time when the number of living rabbits was 199.5. Nevertheless, despite all these over-simplifications, the model correctly predicts some general structural features of population change in predator-prey species. Much progress in natural science is now of this kind. Once we have a successful model, we can try building a little more real-life complexity back into it, step by step, but the models will always be vastly simpler than real life itself – otherwise they would be too complex to analyze.

Sometimes there is no workable alternative to model-building. For example, biologists wonder why two-sex reproduction is the norm for animals, since reproduction by three sexes or none is possible in principle. If you want to understand why a phenomenon *doesn't* occur, you can't go out and observe and measure it. Instead, a good strategy is to build hypothetical models of the phenomenon to see what goes 'wrong' with it. You might study a model where both two-sex reproduction and three-sex or no-sex reproduction occur, to see which does better, perhaps in achieving genetic variation within the species, which enables it to adapt evolutionarily to changes in the environment. Such models aim not to predict observed quantities but to explain an absence.

2. Models in philosophy

Humans are a classic example of messy complex systems. In one way or another, much – though not all – of philosophy is about humans. Thus moral and political philosophy mainly concerns a good human life and a good human society. Philosophy of science concerns human science; philosophy of concerns human art; philosophy of language concerns human language. Though philosophy of art mind pays some attention to non-human animal minds, its main focus is on human minds, and in any case non-human animals are messy complex systems too. Though in principle epistemology concerns all knowledge, in practice it mainly concerns human knowledge. Logic and metaphysics are partial exceptions, since they tend to proceed at a level so fundamental that informative, precise, exceptionless laws are obtainable. For the rest, however, one might expect a model-building strategy to be appropriate.

That isn't how most philosophers have seen it. Many still aim at exceptionless laws, even about messy complex systems – humans – for whom natural scientists have mainly abandoned that ambition. In that respect, philosophers have done their field a disservice, by inadvertently setting it up for failure. People who contrast progress in natural science with deadlock in philosophy often do so on the basis of a false image of natural science. Failing to appreciate how much scientific progress consists in building better models, they fail to ask how much philosophical progress consists in building better models too.

One example of progressive model-building in philosophy is epistemic logic, which advances in just that way. Its models are not universal laws; they involve grossly unrealistic simplifications. Nevertheless, they cast light on human knowledge in the manner of a scientific model.

When philosophers work with probabilities, they typically use models without emphasizing the fact. For instance, a simple model of uncertainty is a lottery. To make things definite, suppose that exactly 100,000 tickets have been sold, numbered in order; there is just one winner, chosen at random. Thus if you have one ticket, its probability of losing is 99,999/100,000. Which statements about the lottery should you accept? You might decide that requiring 100% certainty is unreasonably demanding, and resolve to accept just statements with a probability of at least 95%. Immediately, there is a problem. By your rule, you accept the statement that the winning number is at least 5,001 (because its probability is 95%), and you accept the statement that the winning number is at most 95,000 (because its probability

is 95%), but you refuse to accept the statement that the winning number is between 5,001 and 95,000 inclusive (because its probability is only 90%). Thus you accept each of two statements separately, but you refuse to accept the result of putting them together, their conjunction. A politician who did something like that on television in an election campaign could expect to get crucified. You might think that 95% was a bad choice of threshold for acceptance, and choose a different threshold. But a little calculation shows that the only thresholds for acceptance which avoid such problems, even when there are more tickets, are 0% and 100%. Since a threshold of 0% means accepting every statement whatsoever – total credulity – you are back to a threshold of 100%, the standard of certainty you already rejected as unreasonably demanding. Thus even such a 'toy' model can illustrate the difficulties of basing acceptance and rejection on information about probabilities.

If you think about the lottery model, you can quickly identify some of its simplifying assumptions. For instance, it assumes that you know exactly how many tickets have been sold. In practice, the organization running the lottery may not announce or even know how many tickets have been sold; even if they announce a number, you may give a nonzero probability to the hypothesis that they are mistaken or lying. Then you may also give a nonzero probability to the winning number being 100,001 (since more than 100,000 tickets may have been sold), and you may give a higher probability to the winning number being 1 than to its being 100,000 (since fewer than 100,000 tickets may have been sold). But taking account of all those realistic complications is not time well spent. Thinking about the simple model takes one more quickly to the heart of the problem. When more complex probabilistic models are needed to understand more intricate problems, mathematically-minded epistemologists construct them too.

In philosophy of language, model-building as understood here goes back at least to Rudolf Carnap. An example is his theory of the meaning of *modal operators*, words like 'possibly' and 'necessarily'. He treated them as building up more complex sentences from simpler ones: thus from the sentence 'Everything changes' they make sentences such as 'Possibly everything changes' and 'Necessarily everything changes'.

Logicians had already designed precise formal languages with symbols for logical words such as 'not', 'or', 'and', 'something', and 'everything', which enable one to build up more complex expressions from simpler ones, without limit. We can think of them as taking expressions as input and delivering more complex expressions as output. For example, if you input the sentence 'Everything changes' to 'not', the output is the sentence 'Not everything changes'. Logicians also had a framework for analyzing the meaning of such general words. To each expression, simple or complex, they assigned something called its *extension*, encoding its application to the world. For instance, since the word 'red' applies to red things and not to non-red ones, its extension includes the former and excludes the latter. If a sentence applies to the world, its extension is truth; if it doesn't apply to the world, its extension is falsity. 'Not' makes an output sentence the opposite in extension of the input sentence: if 'Everything changes' is true then 'Not everything changes' is false, while if 'Everything changes' is false then 'Not everything changes' is true. In effect, the extension of 'not' just swaps truth and falsity. 'Or', 'and', 'something', and 'everything' work similarly: they each transform the extension of the input into the extension of the output. Operators which do that are called *extensional*. For each expression of the language, such rules determine its extension from the extensions of the simple words from which it is built up. That helps explain how we can understand complex sentences we never previously encountered by understanding the familiar words of which they are made and how they are put together. In effect, such extensional semantics is an elementary – but very powerful – model of linguistic meaning, though people did not think of it like that at the time.

Carnap wanted to add symbols for 'possibly' and 'necessarily' to the formal language. His problem was that such modal operators don't fit the previous model of meaning: they are not extensional.

For suppose that 'possibly' *is* extensional. Then if I pick a sentence 'X' and don't tell you what it is, but only whether it's true, you should be able to work out whether 'Possibly X' is true. In one case you can do that: if I tell you that 'X' is true, you can work out that 'Possibly X' is true too, since actuality implies possibility. But if I tell you that 'X' is *false*, you can't work out whether 'Possibly X' is true. I haven't given you enough information; the answer depends on what 'X' is. For example, if 'X' is 'Napoleon won at Waterloo' (false), then 'Possibly X' is true: although Napoleon lost, he could have won. But if 'X' is '5 is more than 6' (also false), then 'Possibly X' is false too: 5 could not have been more than 6. Thus the extension of 'X' doesn't always determine the extension of 'Possibly X'. 'Possibly' is not extensional.

Similarly, suppose that 'necessarily' is extensional. Then you should be able to work out whether 'Necessarily X' is true. In one case you can do that: if I tell that 'X' is false, you can work out that 'Necessarily X' is false too, since necessity implies actuality. But if I tell you that 'X' is *true*, you can't work out whether 'Necessarily X' is true. I haven't given you enough information; the answer depends on what 'X' is. For example, if 'X' is 'Napoleon lost at Waterloo' (true), then 'Necessarily X' is false. But if 'X' is '6 is more than 5' (also true), then 'Necessarily X' is true too. Thus the extension of 'X' doesn't always determine the extension of 'Necessarily X'. 'Necessarily' is not extensional.

Carnap solved the problem by considering not just extensions in the *actual* world, the way things are, but profiles of extensions over all *possible* worlds, ways things could have been. He borrowed the idea of possible worlds from Leibniz, though he preferred to use more linguistic entities, 'state-descriptions'. He called the profiles *intensions*. For example, since the extension of 'Napoleon lost at Waterloo' is truth in every world in which Napoleon lost at Waterloo and falsity in every world in which Napoleon did not lose at Waterloo, the intension of 'Napoleon lost at Waterloo' assigns truth to each of the former worlds and falsity to each of the latter ones. Carnap's crucial insight was that although the extension of the input to a modal operator doesn't always determine the extension of the output, the *intension* of the input *does* always determine the intension of the output. He gave rules for calculating the latter in terms of the former. He interpreted 'possibly' as 'in some possible world' and 'necessarily' as 'in every possible world'.

In more detail, Carnap's rule for 'possibly' is that if the input is true in some possible world, then the output is true in *every* possible world, while if the input is false in every possible world, then the output is also false in every possible world. Thus 'Napoleon won at Waterloo' is true in some possible world, so 'Possibly Napoleon won at Waterloo' is true in every possible world. But '5 is more than 6' is false in every possible world, so 'Possibly 5 is more than 6' is also false in every possible world. Thus the intension of the input determines the intension of the output; 'possibly' is intensional rather than extensional.

For 'necessarily', the rule is that if 'X' is true in every possible world, then 'Necessarily X' is also true in every possible world, whereas if 'X' is false in some possible world, then 'Necessarily X' is false in *every* possible world. Thus 'Napoleon lost at Waterloo' is false in some possible world, so 'Necessarily Napoleon lost at Waterloo' is false in every possible world. By contrast, '6 is more than 5' is true in every possible world, so 'Necessarily 6 is more than 5' is also true in every possible world. Thus the intension of the input determines the intension of the output; 'necessarily' is intensional rather than extensional.

Since the rules for extensional operators like 'not', 'or', 'and', 'something', and 'everything' work for extensions in any possible world, Carnap easily adapted them to calculating intensions. The upshot was a complete intensional semantics for his whole formal language: every formula,

however complex, has an intension, determined step by step from the intensions of the simple constituents out of which it is composed. It's a significantly more sophisticated model of meaning than extensional semantics. Through the work of Richard Montague, David Lewis, and many others, Carnap's intensional semantics has massively influenced both philosophy of language and semantics as a branch of linguistics. Although the models have become ever more elaborate, they preserve the crucial move from extensions to intensions.

Carnap worked in a more model-building spirit than his predecessors. He didn't construct his formal language to do mathematics in, or to reveal the hidden essence of all languages. He constructed a simple model language to demonstrate a way for modal operators to work. As we learn ever more of the extraordinary complexity underlying even the most ordinary conversations, philosophers of language and linguists may have to rely increasingly on a model-building methodology.

3. Working models, counterexamples, and error-fragility

Models are fun. You can play with them. That's not just an incidental side benefit; it's what they are for, in both natural science and philosophy. We learn by manipulation, playing about: if you can't manipulate the real thing, a good second-best is often to manipulate a model of it. You can fiddle with this or that component, changing it slightly to see what difference it makes, what varies with what. That way you come to understand more deeply how the model works. If the model is any good, you thereby come to understand better how the real thing works too. For instance, you can't arbitrarily change how English works, to see what difference it makes, but you can arbitrarily change the rules of an artificial language, and calculate the consequences.

To be easily manipulated, a model should be defined in mathematically or logically precise and tractable terms. If the definition is vague, or too complicated, its consequences are unclear: one has to fall back on one's prior philosophical instincts to guess how it behaves, instead of using the model to test those instincts. By contrast, a well-defined model allows one to calculate rigorously how it and variations on it behave, bypassing those prior instincts, and so to learn something unexpected. With a model-building methodology, rigor and playfulness go naturally together.

The rigor of model-building is not the rigor most philosophers are used to. Traditional philosophical rigor requires dismissing a claim once a counterexample to it has been given. In that sense, most models are born refuted, because they involve false simplifying assumptions. For instance, models in epistemic logic typically treat agents as logically perfect. Some philosophers dismiss those models accordingly.

In physics, models of the solar system may treat a planet as a point mass, as if all its mass were concentrated at its center. Of course, physicists know quite well that planets are not point masses and do not behave exactly like them. Nevertheless, physicists do not dismiss such models, for they also know that much can be learned from them. By contrast, if one tried to write into the model a fully accurate description of the planet, with all its craters and bumps, the result would be too complicated to permit calculation. It takes skill to distinguish amongst the features of a model those which have lessons to teach us from those which are mere artefacts of the need to keep things simple. Philosophers are having to learn that skill.

To many philosophers, dismissing the true counterexample rather than the false generalization seems like a disregard for truth. It would indeed be intellectually irresponsible to go on *believing* the generalization in the face of a clear counterexample. But that's not the model-building attitude. One can recognize that a generalization is both false and a key component of a model that points us towards genuine truths.

If counterexamples don't refute a model, what does? Within the model-building methodology, what displaces a model is a better model. Part of its superiority may be that it

deals more adequately with counterexamples to the old model, but it should also reproduce in its own way the old model's successes. A new model with that combination of virtues may be very hard to find.

Model-building contrasts with the methodology of *conjectures and refutations*, championed by Karl Popper. On the crude version of his view, scientists put forward bold conjectures, informative universal generalizations, which can be falsified but can never be verified. A single negative instance, a counterexample, will falsify the generalization; no finite number of positive instances will verify it. Scientists do their utmost to refute it, by finding such a counterexample. Once it is refuted, they put forward another bold conjecture, and so on.

One problem for such a falsificationist methodology, in both natural science and philosophy, is that it is *error-fragile*. In other words, a single mistake can have disastrous consequences. For suppose that we are testing a bold conjecture, and take ourselves to have found a counterexample. As good falsificationists, we dismiss the conjecture and go on to the next one. But what if the counterexample was a mistake? We are fallible; sometimes we misjudge single instances. In that case, the original conjecture may have been true after all. But we never return to it; we are too busy testing new bold conjectures. Philosophers' reliance on counterexamples can be alarmingly close to crude falsificationism: once a counterexample is accepted, there's no going back on it. By contrast, the model-building methodology is much less error-fragile, for it gives no such decisive power to a single judgment. Models are compared over a variety of dimensions.

None of this means that philosophy should go over entirely to a model-building methodology. In some areas, such as logic, we have found many true and informative universal generalizations. In others, good models may be too much to expect. Even where good models are available, as in epistemology, we may do best by using *both* methodologies. For if each independently pulls in the same direction, that's stronger evidence that it's the right direction. Such a combination of methodologies is more robust, unless they pull in opposite directions. The potential of the model-building methodology for philosophy is only beginning to be explored. Its scope and limits should be clearer fifty years from now.

REFERENCES

Carnap, R. (1947). *Meaning and Necessity.* Chicago: University of Chicago Press;

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press;

Williamson, T. (2017). Model-building in philosophy. In R. Blackford and D. Broderick (Eds.) *Philosophy's Future: The Problem of Philosophical Progress* (pp. 159-173). Oxford: Wiley Blackwell;

Williamson, T. (2018). *Doing Philosophy: From Common Curiosity to Logical Reasoning.* Oxford: Oxford University Press.