Brave New Worlds

How computer simulation changes model-based science

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University.

07/2021.

The contents of this thesis are entirely my own work. To the best of my knowledge, all sources I have used and any assistance received in preparing this thesis have been acknowledged.

Word count: 76,000.

Lachlan Douglas Walmsley

21/07/2021

For myself

Acknowledgements

I would like to thank my partner, Courtney, for her love and support, for accepting and joining the philosophy community, and for putting up with me when I have been at my worst. Your excellence has been a source of inspiration.

I would also like to thank my family, especially my mother. Mum taught me to love learning and to love teaching, both of which have served me well during my candidature. In addition to her emotional support, Mum has provided me with financial support. I cannot help but acknowledge the many privileges I have had, which many lack and without which I would not have been able to finish this thesis.

Of course, I must thank Kim, my supervisor. Kim created a philosophy family I am humbled to have been a part of and made the ANU feel like home. He also allowed me to follow my philosophical passions (for better or worse) even though I had funding through his project on the evolution of social complexity. As anyone who reads this thesis will discover, there's not much on that topic in these pages. He will be a friend for life.

Likewise, I must thank John Matthewson. He's a top bloke that really helped me to get a handle on the philosophy of models and to answer some of Kim's toughest questions.

I would like to thank all the friends I have made during grad school. I will resist giving specific names for fear of leaving anyone off the list. They know who they are. And they made being a philosopher for a few years one of the best experiences of my life.

Finally, I must thank Schneeball the sloth, who kept smiling even through the darkest times.

Abstract

A large part of science involves building and investigating models. One key feature of model-based science is that one thing is studied as a means of learning about some rather different thing. How scientists make inferences from a model to the world, then, is a topic of great interest to philosophers of science. An increasing number of models are specified with very complex computer programs. In this thesis, I examine the epistemological issues that arise when scientists use these computer simulation models to learn about the world or to think through their ideas.

I argue that the explosion of computational power over the last several decades has revolutionised model-based science, but that restraint and caution must be exercised in the face of this power. To make my arguments, I focus on two kinds of computer simulation modelling: climate modelling and, in particular, high-fidelity climate models; and agent-based models, which are used to represent populations of interacting agents often in an ecological or social context. Both kinds involve complex model structures and are representative of the beneficial capacities of computer simulation. However, both face epistemic costs that follow from using highly complex model structures. As models increase in size and complexity, it becomes far harder for modellers to understand their models and why they behave the way they do. The value of models is further obscured by their proliferation, and a proliferation of programming languages in which they can be described. If modellers struggle to grasp their models, they can struggle to make good inferences with them. While the climate modelling community has developed much of the infrastructure required to mitigate these epistemic costs, the less mature field of agent-based modelling is still struggling to implement such community standards and infrastructure. I conclude that modellers cannot take full advantage of the representational capacities of computer simulations unless resources are invested into their study that scale proportionately with the models' complexity.

Contents

0

Introduction

A large part of science involves building and investigating things called "models." Lots of different things can be models, from lab rats to mathematical relationships. The key feature of science that uses models, which we can call "model-based science," is that one thing is investigated as a way of learning about some rather different thing. How scientists go about making inferences from a model to the world, then, is a topic of interest to philosophers of science. It should also be of interest to the lay person. When policymakers base their decisions and actions about, say, climate change, on the results of very complex computer programs, what justifies these decisions and actions?

Not all models are computer programs, very complex or otherwise, but an increasing number of them are. In this thesis, I wish to examine the epistemological issues that arise when scientists use computer simulation models to learn about the world or to think through their ideas. For the purposes of my project, I will set aside physical models including model organisms like the lab rat mentioned above, focusing only on models consisting of mathematical relationships. Motivating my examination is the question of what the primary epistemological differences are that arise from the move to modelling with computers. As I see it, there is no deep difference between those mathematical models that are investigated with computer simulation and those that are not. Instead, the difference between computer simulation models and other mathematical models is simply that former are investigated with digital computers and the latter are not. However, even this deflationary way of drawing the distinction has major consequences for the kinds of inferences modellers can make and how to make those inferences well.

The amount and complexity of the mathematical relationships that can be investigated with the assistance of digital computers increases as our technology rapidly advances. Consequently,

modellers can investigate more mathematical structures and represent more real-world processes than they ever could before. Computer simulation, without a doubt, has revolutionised model-based science and vastly increased its scope. The rise of computer simulation, then, is an epistemological gain within model-based science, but this is also why there are distinctive epistemic questions about computer-investigated models.

Bigger is not always better, however. Highly complex structures and high-fidelity representations come with an epistemic cost. As a mathematical structure increases in size and complexity, it becomes much harder for modellers to understand the structure and why it behaves the way it does. If modellers struggle to grasp their models, they can struggle to make good inferences with them. There is a trade-off, then, between how comprehensive a model is as a representation and how comprehensible the model is.

Because of the potential gains in realism and the capacity to represent different systems or familiar systems at different levels of organisation, the epistemic costs that come with more complex models do not imply that these models should be abandoned. Instead, in many cases, the resources put into investigating the model structure as well as ensuring that it represents accurately must be increased and scaled proportionally to the increased complexity of the model structure. In fact, as I will argue, the increased complexity of model structures, the number of new mathematical structures it is possible to investigate, the number of programming languages with which these mathematical structures may be described, and the complicated relationship between models and empirical data, mean that modelling communities must take care to manage the construction and analysis of models. In some cases, this requires the implementation of standards to avoid a state of anarchy where models proliferate but the epistemic value of any one model is completely unknown. This is certainly true in one domain of computer simulation, known as agent-based modelling, where many different programming languages have been used to describe many models, which are often poorly communicated within journals, leading to suspicion of these models compared with more traditional approaches.

The main take-home message of my examination is that the explosion of computational power over the last several decades has revolutionised model-based science, but that restraint and caution must be exercised in the face of this power. This includes the use of multiple models of differing levels of complexity, but also the use of frameworks that allow for useful comparisons and connections to be drawn between these different models. Just as the value of more complex models is limited if those models go unchecked and do not perform as desired, having more models adds limited value to a modelling community if those models cannot be used together. This is closely related to the practice of robustness analysis, familiar to philosophers of science and many modellers, in which multiple different models are used to investigate the stability of a result across models and its susceptibility to different assumptions used in model construction. However, many current discussions of robustness analysis omit details about how, in practice, modellers can successfully and systematically compare their models—an omission my thesis seeks to address.

To make my arguments, I will focus on two kinds of computer simulation modelling: climate modelling and, in particular, high-fidelity climate models; and agent-based models, which are used to represent populations of interacting agents often in an ecological or social context. Both kinds involve complex model structures than can only be investigated with computer simulation and are representative of the beneficial capacities of computer simulation. However, both face challenges in handling the consequent epistemic costs that follow from using these complex structures. The two groups make for a useful comparison, as the climate modelling community has developed much of the infrastructure required to mitigate these epistemic costs, while the less mature field of agent-based modelling is still struggling to implement such community standards and infrastructure. Yet both cases indicate the need for frameworks that permit the systematic comparison of models to enable computer simulation modelling to fully deliver the benefits of its greater representational capacities.

So far, I have spoken about models without defining them in any detail or even providing many examples. In Chapter 1, then, I describe three models of varying complexity. Afterwards, I will present a framework for analysing scientific models and computer simulations, which I will use throughout this thesis. The framework I adopt is a member of the similarity views of Ronald

Giere, Michael Weisberg, and Peter Godfrey-Smith. According to this view, scientific models are structures, which are abstract or physical entities that can be objects of study all of their own, there is a model description which specifies the structure, and a target, to which the structure can be compared. I will also use this framework to define computer simulation as involving a model with an abstract structure that also has a model description which is manipulated by a digital computer.

Chapter 2 makes a claim about the epistemology of computer simulation: computer simulation enables the investigation of abstract model structures that are more complex than ever before, but this complexity comes with its own set of epistemic challenges. This chapter also contributes to the philosophy of modelling and the literature on modelling trade-offs, arguing that the complexity of model structures, rather than the causal heterogeneity among targets, is sufficient to produce modelling trade-offs.

In Chapter 3, I take a close look at the concept of robustness analysis, a concept which is important in any discussion of the epistemology of models. I argue that robustness analysis and scientific understanding are closely connected, with robustness analysis being a method for examining causal dependencies and scientific understanding being achieved when a scientist has grasped causal dependencies. This chapter contributes to the literature on robustness analysis, both in the philosophy of climate science and the philosophy of science more broadly. It does so by clarifying that robustness analysis should be conceived as a process directed toward fostering understanding and by demonstrating a practical method by which it achieves this aim. This can be contrasted with the view that robustness analysis is a method by which the results of models can be confirmed, at least to some minor degree.

In Chapter 4, I examine the representational capacities of agent-based models and investigate the kinds of explanations they can support. Many agent-based modellers describe agent-based models as supporting mechanistic explanations because these models explicitly represent the interacting components underlying phenomena. However, both agent-based models and their targets are distinctly

non-machine-like because they are populations. This presents a puzzle. In this chapter, I explore the relationship between mechanistic explanation and agent-based models by describing three ways in which population phenomena can be represented and explained. The third way, involving the use of agent-based models, has distinctly mechanistic features, including the decomposition of systems into their lower-level entities and a focus on the activities of those entities. However, there are also important departures from the mechanistic framework. My analysis results in a partial vindication of claims regarding the mechanistic capacities of agent-based models and suggests more work should be done on the relationship between population thinking and explanation. In making my argument, I contribute to existing discussion on the relationship between mechanistic explanation and population phenomena.

In Chapter 5, I examine the relationships between model description and model structure, comparing equation-based models to agent-based models. I argue that using computer simulation and model descriptions that can be manipulated by digital computers leads to uncertainty regarding this relationship. In the case of equation-based models, digital implementation requires that continuous mathematics is discretised, leaving some uncertainty about whether the structure investigated through simulation is relevantly similar to the structure which the modellers desire and intend to investigate. Agent-based models, at least initially, may appear to avoid this uncertainty because they typically do not involve the discretisation of continuous mathematics. However, as I argue, there remains a non-trivial degree of uncertainty regarding whether the description investigated through simulation specifies the structure intended. This chapter contributes to the philosophy of models and computer simulation by showing that the distinction between model structure and model description is particularly important in understanding the epistemology of computer simulation as compared to non-simulation models, where the relationship between description and model is more transparent due to the relative simplicity of both.

In Chapter 6, I describe how the epistemic challenges facing models with highly complex structures described in Chapter 2 can be addressed if, as established in the Chapter 3, robustness analysis is not the answer. I argue that thoroughly investigating highly complex simulation models

requires a proportionally large amount of community coordination. While examples of such community coordination can be found in climate science, they are mostly absent in the context of agent-based modelling. My arguments also reveal that heterogeneity among target systems creates a further problem for complex models. In addition to hampering the generalisability of complex and detailed models, target heterogeneity divides resources, prohibiting research communities from establishing the practices that are required to ensure that complex models perform as well as possible. Consequently, fields characterised by target heterogeneity are likely to have less success with high fidelity modelling than fields focused on a singular target or small set of similar targets. This chapter contributes to the epistemology of computer simulation by describing the scientific practices that improve the chances of producing good results with highly complex models.

Now, let's get into it.

1

What is a scientific model?

In this chapter, I will describe three models of varying complexity. I will then present a framework for analysing scientific models and computer simulations, which I will use throughout this thesis. The framework I adopt is a member of the similarity views of Ronald Giere, Michael Weisberg, and Peter Godfrey-Smith. According to this view, scientific models are structures, which are abstract or physical entities that can be objects of study all of their own, a model description which specifies the structure, and a target, to which the structure can be compared. I will also use this framework to define computer simulation as a model with an abstract structure that also has a model description which is manipulated by a digital computer.

1.1 Introduction

Nelson Goodman (1976) was not exaggerating when he said "model" was a promiscuous term (c.f. Downes, 2011, p. 757). Various models that you may have heard mentioned include, "mathematical models," "computational models," "statistical models," "scale models," "toy models," "equilibrium models," "dynamical models," "mental models," "fashion models," and so on. Since Goodman wrote, "model" has not become chaste. Indeed, there are more different things called "models" today than there were in the 70s!

Such is the diversity among scientific models—here at least we can bracket off the fashion and Instagram models—that some philosophers have argued that the general category of scientific model resists general analysis at all (e.g. Downes, 1992; O'Connor & Weatherall, 2016; Veit, 2020). If they are correct, then anyone pursuing a project in which they attempt to say something about the

epistemology of computer simulation is unlikely to get far and is unlikely to say something interesting. But this is the very project *I* am pursuing.

The aim of this chapter, then, is to state clearly what I mean by "scientific model" and "computer simulation model" as well as to go some way toward convincing you that my analysis and the sorts of conclusions I will draw throughout this thesis are not so general as to be vacuous. I start, in section 1.2, by presenting a very small snapshot of scientific models in all their diversity. I look at three scientific models. The first two are climate models that are representative of the poles of a spectrum running from the simplest mathematical models, which can be manipulated with "back of the envelope" reasoning, to the most complex climate models that can only be investigated through the use of incredibly powerful digital computers. The third example is another computer simulation model, but one that can be investigated with a computer no more powerful than a good laptop. These exemplars have been chosen both because they make for a nice sample of a spectrum of complexity and because they are different *kinds* of model: the first is an equilibrium model, the second is an equation-based dynamical model, and the third is an agent-based dynamical model. We will not dive into the differences between these kinds just yet, but let it be known that these exemplars have been chosen to represent a fair spread of diversity relative to my philosophical aims of exploring the epistemology of computer simulation.

In section 1.3, I describe a framework that will assist me when analysing my case studies and in generalising the lessons of this analysis to modelling and computer simulation more broadly. This framework is based on Michael Weisberg's *interpreted structures* view, itself a member of a family of similar views held by Ronald Giere and Peter Godfrey-Smith. I adopt and focus only on what I take to be the core aspects of this view, which I call the *simple similarity* view. This view focuses on modelling as involving a model structure (the thing studied), a model target (the thing about which the model structure is used to make inferences), and a model description (the thing that specifies the model structure). This framework is abstract enough to apply to scientific models in all their diverse forms while still highlighting some of the ways in which modelling can go well or poorly and, in particular, how inferences made with models change when computer simulation gets involved.

Without further delay, let's look at some models.

1.2 Three exemplar models

1.2.1 Energy balance models

Energy balance models (EBMs), are global climate models that are built with the intention of being very simple. The simplest is a *zero-dimensional* EBM (McGuffie & Henderson-Sellers, 2013, Chapter 3). In a 0D EBM, we determine the Earth's temperature by considering incoming and outgoing radiation. This is shown schematically in figure 1.1. The incoming radiation contacts the Earth only from one direction, making the incoming solar radiation the solar constant S multiplied by the area of a circle πR^2. The zero-dimensional EBM is a simplification of one-dimensional EBMS, which represent a planet's latitude bands. In a 0D EBM, the Earth is represented as a uniform black body, which radiates energy at the rate of the Stefan-Boltzman constant σ. Since the Earth is a sphere and the area of a sphere is $4\pi R^2$, the outgoing infrared radiation is $4\pi R^2 \sigma T^4$, where T is the temperature. This simple system will be in equilibrium when

$$\pi R^2 S = 4\pi R^2 \sigma T^4 \tag{1.1}$$

or, simplifying:

$$S/4 = \sigma T^4. \tag{1.2}$$

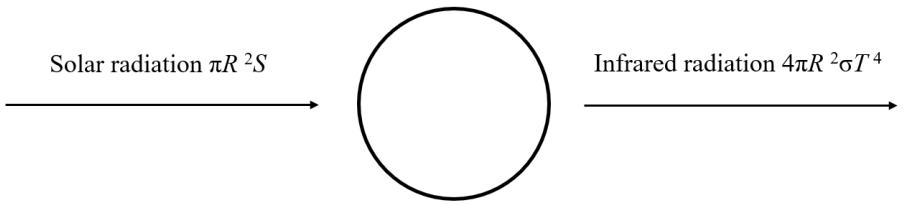

Figure 1.1 The Earth's temperature in this highly idealised energy balance model is determined by the incoming solar radiation and the outgoing infrared radiation.

This system is a little too simple to be informative, but with only an extra two mathematical relationships—the incoming and outgoing radiation—we can produce something of use. The next relationships we will add to our representation are the albedo effect and the greenhouse effect. The albedo effect inhibits the incoming solar radiation by reflecting shortwave radiation back out into space. The biggest contributors to the Earth's albedo are large white surfaces, which are highly reflective, including clouds, ice sheets, snow, and so on. The incoming radiation is now $S/4$ multiplied by $(1 - \alpha)$, where α is the albedo factor. So, if $\alpha = 0$, then no incoming solar radiation is reflected and if $\alpha = 1$, then all the incoming solar radiation is reflected.

The greenhouse effect describes the impact of gases like CO_2, fluids like H_2O, and particles like dust, on the atmosphere's transmissivity—that is, how well infrared radiation passes through it. Just as the albedo inhibits the incoming solar radiation, the greenhouse effect inhibits the outgoing infrared radiation. The outgoing radiation is now σT^4 multiplied by a transmissivity factor ε, where the higher the greenhouse effect, the lower the transmissivity because less radiation is transmitted through the atmosphere. So, if $\varepsilon - 1$, then all the outgoing infrared radiation is transmitted through the atmosphere, and if $\varepsilon = 0$, then no outgoing infrared energy is transmitted through the atmosphere.

This slightly richer, though still very simple, system is shown schematically in figure 1.2 and will be in equilibrium when

$$(1 - \alpha)S/4 = \varepsilon\sigma T^4 \tag{1.3}$$

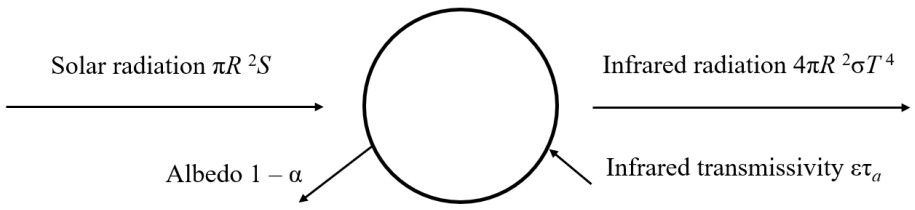

Figure 1.2 The Earth's temperature in this slightly less idealised energy balance model is determined by the incoming solar radiation adjusted for the albedo effect and the outgoing infrared radiation adjusted for the infrared transmissivity of the atmosphere.

0D EBMs have a couple of uses. First, they can be useful for making illustrative points. For example, we can see that if we increase the albedo factor from its equilibrium value to a larger value α_1, then outgoing infrared radiation will be larger than the incoming solar radiation and the planet will cool:

$$(1 - \alpha_1)S/4 < \varepsilon\sigma T^4 \tag{1.4}$$

Likewise, if greenhouse gas concentration increases and lowers the atmosphere's transmissivity to a new value ε_1, then the incoming solar radiation will be greater than the outgoing infrared radiation and the planet will warm:

$$(1 - \alpha)S/4 > \varepsilon_1\sigma T^4 \tag{1.5}$$

Although obvious, these basic relationships can assist us in producing hypotheses and explanations about more complex climate phenomena. Figure 1.3 shows average global temperatures from 1880 to about 2020. Although the warming trend is clear, there appears to be a multi-decade hiatus running from about 1940 to 1980. A possible explanation for this hiatus (if it really occurred at all[1]) is that, during this time, an increased albedo effect dampened the effects of increasing greenhouse gas concentration in the atmosphere (Edwards, 2010). The increased albedo effect may have been caused by an interaction between chlorofluorocarbons (CFCs) and clouds, which account for about 70% of the Earth's albedo. CFCs can cause smaller water droplets to form, which are more reflective, effectively increasing clouds' albedo factor. Indeed, the "alarm" about an impending ice-age that climate change denialists allege swept through climate science during the 70s[2] had its roots in a

[1] One reason for thinking there is no real hiatus is that climate is typically thought of as a multi-decade phenomenon, with a single unit of climate time being about thirty years. If the graph shown in figure 1.3 is represented as data points averaging temperatures across thirty-year intervals, the apparent hiatus disappears.

[2] In fact, the scientists anticipating global cooling were in the minority. Most papers on global change published during the 1970s suspected that the change in transmissivity would counterbalance and outpace the change in albedo leading to a warmer world and that these cooling projections followed from an over-estimated impact of aerosols on the albedo effect.

concern that (human) aerosol emissions would increase the albedo effect to the point where human action would significantly cool the Earth's climate (Peterson, Connolley, & Fleck, 2008; Rasool & Schneider, 1971). So, by isolating key relationships determining global temperatures, the 0D EBM prompts a hypothesis: *if warming has slowed, could anything have caused an increase in the albedo factor?* It even suggests interventions we could make on the system: *Could we slow warming by painting every rooftop and building white to increase albedo?*

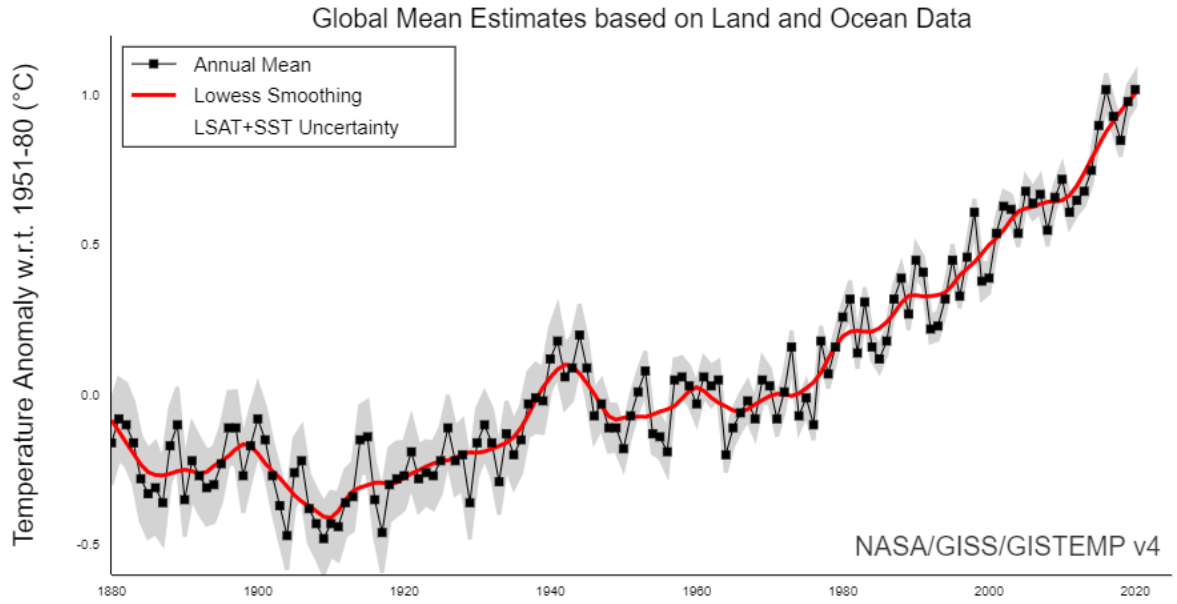

Figure 1.3 Global average temperatures from 1880 to about 2020. The horizontal line represents the average temperature during the warming hiatus from about 1940 to about 1980 (reproduced from NASA Goddard Institute for Space Studies, 2021).

Despite being very simple, the 0D EBM can also be used to make specific predictions about changes in global temperatures by setting the values of the parameters based on empirical data (McGuffie & Henderson-Sellers, 2013). Alternatively, and in virtue of its simplicity, the 0D EBM is also a highly general model, applying to the global climates of many planets. This is because the factors included in the model—incoming solar radiation adjusted for albedo and out-going infrared radiation adjusted for the atmosphere's infrared transmissivity—play a large role in determining the temperatures not only on Earth, but on Mars, Venus, and just about everywhere else. When looking at these other targets, we can make use of the illustrative and explanatory powers of the basic

relationships included in the model: *Why is Venus so hot?* Because transmissivity is very low. *Why is Mars so cold?* Because transmissivity is very high.

To summarise, 0D EBMs are very simple models that encode a few key relationships that determine global temperatures: the incoming solar radiation, adjusted for the albedo, and the outgoing infrared radiation, adjusted for the transmissivity of the atmosphere. Focusing on these key relationships can assist us in thinking about more complex, real-world systems, and using empirical data to instantiate the model's parameters can produce surprisingly accurate estimations of global temperatures and changes in global temperatures in multiple targets. Now, let us turn from the simplest climate models to some of the most complex.

1.2.2 Earth system models

Earth systems models (ESMs) are incredibly complex computer programs, typically requiring multi-person teams to construct and study. Consequently, a close look at the formal descriptions of these models is impossible here; however, I can describe their basic anatomy. The primary component of ESMs and their slightly less detailed counterparts, global circulation models (GCMs), is the *dynamical core*. The dynamical core numerically approximates the so-called "primitive equations." The primitive equations are a set of seven non-linear partial differential equations representing atmospheric flow that were published by Norwegian physicist Vilhelm Bjerknes in 1904 (Bjerknes, 1904; Edwards, 2010; Washington & Parkinson, 2005, p. 49). Picked from the laws of fluid dynamics, these equations included "Newton's laws of motion, the hydrodynamic state equation, mass conservation, and the thermodynamic state equation" (Edwards, 2010, p. 85). By adding another equation for salinity, the primitive equations can be modified to represent the oceans.

Unfortunately, scientists would not be able to use the primitive equations for about another half century because they are mathematically intractable—that is, they cannot be solved.[3] In 1922, English mathematician Lewis Fry Richardson developed mathematical techniques for numerically approximating the primitive equations, turning the differential equations into difference equations. Unlike differential equations, a large set of difference equations can yield a result if only the calculator (human or machine) has enough time. Using observations taken on International Balloon Day of 1910, a day for which the observational record was unusually good, Richardson attempted to produce a weather forecast (really a retrodiction given that Richardson's data was from over a decade earlier) (Edwards, 2010). It took Richardson six weeks to produce his six-hour forecast, which, owing to a calculation error, was terrifically inaccurate.

With the assistance of powerful digital supercomputers, the solutions of Bjerknes' primitive equations can be numerically approximated in a relatively timely fashion and, hopefully, with fewer errors than Richardson's calculations. ESMs still bear some similarities with Richardson's model. Richardson proceeded by carving Europe into a three-dimensional grid of 90 cells and prepared 23 programs for turning observed data into a forecast. While some climate models remain regional, ESMs and GCMs use a 3D grid that represents the atmosphere, the oceans, and the land of the entire planet. Using this grid, the computer steps through discretised versions of the governing equations. Grid sizes vary, from coarser grids with cells spanning 200km squared along the Earth's surface, to finer grids with cells spanning 20km squared (Neelin, 2010, p. 150). Generally, the finer the resolution the better, but finer resolution comes at the cost of computational tractability—that is, the time it takes for the

[3] When a system of partial differential equations has an analytic or closed-form solution, we will have an algebraic expression wherein the dependent variables of the equation can be expressed in terms of the independent variables. These expressions are valuable as they "can represent the behaviour of a general class of systems" and because "various functional dependencies and patterns of behaviour can easily be read off from a closed-form solution" (Winsberg, 2010, p. 33). This is in contrast to, say, numerical solutions, where the results are generated by inputting particular values for the variables, producing results that are not general and that do not have the benefit of being easily interpretable.

computer to complete all the necessary calculations—which places a cap on how closely the primitive equations can be approximated.

While a significant part of the Earth system—the atmosphere and the ocean—can be represented as circulating fluids, the Earth system has many processes that make a difference to climate behaviour that cannot be represented using fluid dynamics alone, such as sea ice and vegetation. Consequently, the other component of an ESM is the *model physics*, which represents these additional processes of the system. Increases in computing power have enabled modellers to expand the model physics, explicitly representing more components of the climate and Earth systems over time, starting with the atmosphere, land surface, and oceans in the 1960s, through to a host of factors and processes including sea ice, aerosols, vegetation, ice sheets, and biogeochemical cycles in contemporary models (Washington, Buja, & Craig, 2009).

ESMs use separate mathematical structures to represent important processes that occur at scales too small to be resolved within an ESM's finite difference grid. These are called *parameterisations*. Shallow cloud formation in the lower atmosphere, for example, is a major contributor to the Earth's albedo—that is, its capacity to reflect incoming solar radiation—and consequently to climate behaviour. However, cloud formation occurs on a scale finer than even the highest resolution grids on the market. Consequently, it must be represented with a parameterisation scheme.

All the model components of an ESM—essentially a collection of smaller models—must be connected together with a coupler. A coupler facilitates data transfer between model components, such as the atmosphere module and the ocean module. Although the coupler enables the representation of coupled dynamics between these components, the coupler itself not really a representational component of the model; there is no structure in the climate system that corresponds to the coupler. Instead, it is an example of what Winsberg (2010) refers to as "hand-shaking algorithms," which are parts of a simulation program whose primary purpose is to connect model components. In the climate case, the coupler also defines the basic architecture of the model and is the part of the simulation

program through which all the components are controlled. Different modelling groups couple their components together in different ways. Alexander and Easterbrook (2015) analysed the architecture of eight models and found two basic architectures: a star-shaped architecture found in US models and a two-sided architecture found in European models. In the case of star-shaped architectures, all components are typically separate and connected together through the coupled, while two-sided architectures see components nested within or connected to the ocean and atmosphere components, which are then coupled together (see Alexander & Easterbrook, 2015, pp. 1225–1226 for a set of helpful diagrams).

Unlike the 0D EBM described above, the ESMs I describe here are not a single model but a family of models. While the 0D EBM consists of the relationships described by equation (1.3), different modelling groups use different pieces of code to specify slightly different mathematical relationships, all of which are intended to represent the Earth system. This is true even when we allow for the fact that we can instantiate the parameters of the 0D EBM in many different ways: in the case of ESMs, there are different models consisting of different mathematical relationships, which can then, in turn, have their parameters differently instantiated. This means that when we talk about ESMs, there is no single piece of code or set of mathematical relationships that we are talking about in the same way there is when we talk about the 0D EBM; the difference between star-shaped and two-sided architectures demonstrates this clearly. Rather, we are talking about a class of models, just as we are talking about a class of cars when we speak of sedans or sports cars.

The reason I am using ESMs as exemplars is because they are paradigmatic examples of complex computer simulation models. If I am to say something about the epistemology of computer simulations in science, then ESMs are an excellent test case in virtue of their immense complexity. Not only do these models consist of a great many more mathematical relationships than a model like the 0D EBM, but they are more complex than many other computer simulations. ESMs are, if you like, a kind of limit case of high-fidelity computer simulation modelling. A further, no less important reason to use ESMs in my analysis is the role that these models play in informing policy makers about anthropogenic climate change. Clarity about how and what kind of knowledge scientists generate with

their models is something that we should all value given the impact that these models have on our lives via policy. Investigating their epistemology as climate models—rather than investigating them solely as instances of computer simulations—is a valuable philosophical endeavour in its own right.

Now, let's turn to the final exemplar to be presented in this chapter.

1.2.3 Physarum polycephalum

My final exemplar is an agent-based model of *Physarum polycephalum*, a multinucleate but unicellular amoeboid organism commonly known as a slime mould. The slime mould may be less familiar to you than the climate, so let me start by saying a few words about this fascinating creature.

In the most active part of its life cycle, known as the "plasmodium stage," the slime mould grows out into a yellow-green blob, searching for sources of food, such as microorganisms and fungus, which it envelops and digests. From there, the slime mould will start to fruit and reproduce. I will follow others in referring to the slime mould in its plasmodium stage as "Physarum" (Jones, 2015).

Physarum is composed of two main materials: a spongy ectoplasm (Gel), as well as a watery endoplasm (Sol). Gel is a kind of spongy muscular tissue—a matrix of actin-myosin fibres—and the watery Sol carries nutrients around the plasmodium, feeding the Gel tissue. In nutrient-rich conditions, the slime mould grows out radially from the nutrient source in a pulsating fashion. A plasmodium sheet is left behind the growth front, which then breaks down to form nutrient networks. Physarum's growth and movement are caused by internal pulses, which emerge as patches of actin-myosin fibres contract and relax. Contractions in the centre of the plasmodium push the Sol outward, causing the slime mould to expand. Areas of the plasmodium that are more often contracted receive less exposure to the life-giving Sol, so die off, leaving only the most efficient paths behind.

Although the slime mould has long been a target of study in cell biology, its many talents have recently gained the attention of behavioural ecologists (Beekman & Latty, 2015; Vogel et al., 2015) and computer scientists (Jones, 2015). For example, Physarum is able to anticipate salient

environmental patterns by embodying external regularities within its internal contraction patterns (Ball, 2008; Pershin, La Fontaine, & Di Ventra, 2008; Saigusa, 2008). The sun rising and flooding a patch with harmful sunlight will cause the slime to recede from that area. But once the light is gone, the slime begins to move back into that area and recedes once more the next day when the light returns. Once Physarum entrains to this oscillation, it may begin to recede from the dangerous patch even before the light is present. Physarum can also be habituated to repellents, coming to ignore them if they are presented repeatedly and in harmless quantities (Boisseau, Vogel, & Dussutour, 2016). Physarum also alters its environment as it moves, leaving a trail of extracellular slime. This acts like a trail of breadcrumbs with which Physarum can track previously explored areas (Reid, Beekman, Latty, & Dussutour, 2013; Reid, Latty, Dussutour, & Beekman, 2012). Despite its simplicity, Physarum is an impressive problem solver, building nutrient networks that can balance distance, robustness, and cost, about as well as human-designed networks (Tero et al., 2010; Zhang et al., 2015). Its networks can also match the solutions to classic geometrical computational problems, like the Towers of Hanoi problem (Reid & Beekman, 2013).[4]

Physarum's ability to problem solve with few if any of the internal resources that we traditionally associate with intelligence and flexible behaviour has led many to attempt to model its behaviour, following its demonstration of doing a lot with a little. We can distinguish between two different approaches toward modelling the slime mould (Gao et al., 2018). Physarum-based modelling aims to describe and explain the slime mould's features and the key causal structures or processes that drive its behaviour. For example, a model might be used to describe how the slime mould builds its

[4] The tower of Hanoi problem involves three pegs and a series of disks of decreasing size stacked one on top of one another. The disks have holes in their centres so they can be placed onto the pegs. The game starts with the disks on one peg, stacked from largest at the bottom to smallest at the top. The aim of the game is to move the entire stack (or tower) to another peg by moving one disk at a time, only ever placing the disks onto other pegs and never placing a disk on a smaller disk. Since it would be too demanding to expect the slime mould to manipulate disks and pegs, (Reid & Beekman, 2013) recreate the problem in two-dimensional space, using a maze that represents the choice-points of the puzzle, with the different routes representing different solutions.

highly efficient networks. Physarum-based computing, on the other hand, takes these insights and exploits them to solve problems. Consider the travelling salesman problem (Jones, 2015, Chapter 9). In the travelling salesman problem, an imagined salesman must visit multiple different locations and end in their original location, taking the shortest path, and visiting each location only once (except the start-end location). The problem is interesting because the number of possible paths increases rapidly as the number of locations increases, meaning that a brute force calculation of all the possible paths and selection of the shortest path is intractable for many instances of the problem. A practical instance of this problem is the optimal storage and retrieval of goods in a warehouse. One could imagine, for instance, a company like Amazon desiring an algorithm that they could use to direct employees or robots around their warehouses more efficiently; Physarum-based computing could assist in the development of such an algorithm.

Jeff Jones (2015) built a series of discrete agent-based models to investigate some of the causal mechanisms underlying Physarum's behaviour. Jones' virtual plasmodium is intended to straddle both Physarum-based modelling and Physarum-based computing by exemplifying the organisational properties of the slime mould and using them to try and solve computational problems. Although his aim is partly pragmatic, Jones is explicit about his concern for the representational fidelity of his model. For him, there are multiple levels of organisation at which we could represent the slime mould: "*atomic-molecular-chemical-actomyosin-plasmalemma-plasmodium*" (Jones, 2015, p. 32). If we represent the slime at a higher level, says Jones, we risk making too many assumptions about the phenomenon to be explained. Going too low, on the other hand, requires making too many empirical assumptions about the underlying causal mechanisms which remain unsettled and would also impose computational costs. Given this, Jones opts for a "particle based reaction-diffusion mechanism behaving as a collective virtual material" (2015, p. 33).[5]

[5] By "particle based," Jones means that he is using an agent-based model—"particles" being another name for "agents" in this context. A "reaction-diffusion mechanism" here refers to a layer into which particles deposit virtual chemoattractant, which then spreads, or diffuses, across the layer. Other agents or particles also react to

Though Jones' model is far more complicated than the 0D EBM, consisting of far more mathematical relationships, it is much simpler than the ESMs, and simulations of the model can be run on a good laptop rather than a supercomputer. Given this, I will provide more explicit details about the model's structure than I did for the ESMs.

Particle-like agents constitute the virtual plasmodium and interact with and within a 2D environment. The agents represent hypothetical units of Gel/Sol interaction. Their position represents the position of Gel—the "relatively stiff sponge-like matrix composed of actin-myosin fibres" (2015, p. 37)—and their movement represents the movement of the protoplasmic solution through the Gel matrix. Figure 1.4 below compares two real Physarum to Jones' virtual plasmodium. This static shot shows the positions of Jones' agents and, therefore, the Gel structure which forms a similar branching pattern to that of the real plasmodium. When the model is in motion, particles can be seen and measured to trace a path back and forth as they slowly flow out to the high concentration areas, resembling the shuttling streaming seen within Physarum under the microscope, in which the nutrient-bearing Sol flows back and forth as the slime contracts and expands.

---

the concentration of virtual chemoattractant in the field. Jones notes that his model differs from classical reaction-diffusion models because those typically use an activator and inhibitor (think food and poison), while Jones' model only uses an activator. Finally, the reference to a virtual material is used by Jones to clarify his level of analysis. That is, he does not intend the processes below the level of the collective to be representational.

Figure 1.4 (a) and (b) show plasmodium inoculated at the centre line, with high nutrient oatmeal agar to the left and plain agar to the right. (c) shows the model plasmodium inoculated at the centre of a virtual dish with high concentrations of chemoattractant to the left and low concentrations to the right (reproduced with permission from Jones, 2015, p. 71).

The Gel/Sol particles produce Physarum's emergent behaviour, such as oscillation and network formation, through their collective behaviour, as shown by figure 1.5. Agents move in the direction they are facing, which is determined either by random, as on the first simulation time-step, or through the sensing procedure just described. If the agent is moving into an empty site on the lattice, then they deposit chemoattractant into the lattice as they move. Agents respond to occupied sites in one of two ways. In the basic, non-oscillatory condition, the agent will stay in place and reorient randomly. This condition, however, does not produce Physarum's characteristic "surging movement." In the organism, resistance can build up from the contracting actomyosin matrix which blocks the movement of Sol around the Physarum's body. In the model, this is approximated in the oscillatory condition, where agents respond to occupied sites by staying put and updating an internal coordinate counter in the direction of their current heading until they find an unoccupied site, moving there and by-passing the occupied sites in between. As Jones states, "This simple mechanism results in temporary blockages until the particles are able to 'push past' each other and results in the emergence of surging movement" (p. 39).

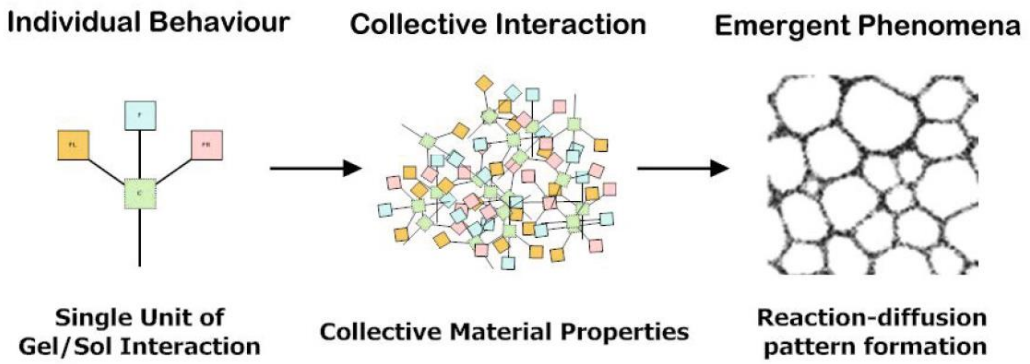

Figure 1.5 These three diagrams show the agents of Jones' model at different resolution. On the left, the individual agents or particles represent gel/sol units, shown here with the front-left, front, and front right sensors protruding from a central node. These sensors are used to coordinate the agent's behaviour. In the centre, the agents are shown clustered together, collectively forming the plasmodium material. And, on the right, we see the population-level tubular structures that emerge from the collective interactions of the individuals (reproduced with permission from Jones, 2015, p. 36).

The virtual plasmodium also grows and shrinks as follows (Jones, 2015, p. 40). First, each particle must be labelled as one meeting the conditions for growth or shrinkage. To do this, each particle checks its local neighbourhood and counts its neighbours within a specified range.[6] If that

[6] Neighbourhoods in spatially implement agent-based models like Jones' typically take two forms: Von Neumann neighbourhoods and Moore neighbourhoods. If you consider an agent to occupy a cell on a two-dimensional grid of square cells, then Von Neumann neighbourhood consists of the cells to the north, south, east, and west of the focal cell, while a Moore neighbourhood consists of these plus the cells to the north-east and north-west and south-east and south-west. According to Uri Wilensky and William Rand state in their textbook on agent-based modelling, "In general, a Moore neighbourhood gives you a better approximation to movement in a place, and since many [agent-based models] model phenomenon where planar movement is common, it is often the preferred modelling choice for discrete motion" (2007, p. 236). Jones' model uses a Moore neighbourhood, though typically expanded out beyond the most basic 3x3 Moore neighbourhood. For example, in the growth procedure, a 9x9 neighbourhood is used, while a 5x5 neighbourhood is used in the shrinkage procedure (2015, p. 132).

number falls between parameters *Gmin* and *Gmax,* specifying the minimum and maximum number of neighbours a particle needs to grow, then that particle also generates a random number between 0 and 1. If the result is smaller than another parameter *pDiv,* then the particle gets tagged for growth. If the number of neighbours the particle has falls between *Smin* and *Smax,* then the particle will get tagged for shrinkage. At regular intervals, the scheduler—the bit of the program responsible for executing the model's computational instructions—applies the rules for growth or shrinkage to the tagged particles. Roughly, to grow, a particle checks a random adjacent cell and, if it's empty, a new particle is created in that cell. To shrink, the particle is simply removed.

These simple movement, growth, and shrinkage rules are enough to allow Jones' particles to collectively reproduce many of Physarum's behaviours, including oscillation, movement and network formation. For example, a random distribution of particles will spontaneously produce a network of agents that exhibit shuttle streaming, flowing back and forth just like Sol does within real plasmodium. Network formation is based on a positive feedback loop as agents are drawn to chemoattractant and deposit chemoattractant into the lattice when moving. Agents in locations with less chemoattractant quickly move to regions with more chemoattractant further reducing concentrations of chemoattractant in these areas.

My reason for using Jones' model as an exemplar are as follows. First, it is, like the ESMs, a model that is too complicated to be investigated without computer simulation. However, it is much simpler than the ESMs, so it provides a nice middle point between the ultra-simple 0D EBM and the ultra-complex ESMs. Perhaps more importantly, however, Jones' Physarum model is different in kind to ESMs, not just in degree. For the most part, ESMs are equation-based models, which numerically approximate some set of mathematically intractable differential equations that govern the global behaviour of the system. For example, the governing equations mentioned in 1.2.2 representing atmospheric flow. Jones' model, on the other hand, is an agent-based model.[7] For such a model, there

[7] I must note now that there are multiple names used for these sorts of models, including multi-agent models, individual-based models, and agent-based models. I will be using the terms interchangeably; I'm a lumper, not a

are no mathematical relationships governing the system's behaviour at a global level. Instead, the system's global behaviour—in the case of the Physarum model, this global behaviour includes the networks that the slime mould creates—is generated by a population of interacting units, and it is the unit behaviour that is governed by the mathematical rules. As I will describe later in this thesis, agent-based models face some epistemic problems that are not faced by equation-based models, so including this diversity in my analysis is important for saying something of value about the epistemology of computer simulation. Indeed, as in the case of ESMs, agent-based models can be used to inform policy, so understanding the epistemology of agent-based models should be of interest even to those beyond the philosophy of computer simulation and even philosophy of science altogether.

The exemplars described throughout section 1.2 represent just a small slice of the impressive diversity among scientific models and computer simulations. In order to make general statements about the epistemology of computer simulation, we need a framework that is abstract enough to accommodate these different cases and many more, but that also isolates factors that are important enough that these generalisations do not veer into vacuity. It is to that framework that I turn now.

1.3 The simple similarity view

Let me begin by restating that my aim here is not to develop and defend *the right and complete* taxonomy or ontology of scientific models. Instead, my aim is to describe a framework that can help me shine my spotlight on patches of model diversity and to compare and analyse the models and practices of some modellers and simulationists within these patches. There may be cases that do not fit neatly or at all into the framework. This is okay. As the saying goes, "all models are wrong, but some are useful" (Box, 1979, p. 2), and I see my view of models as a kind of model itself: a simplified and

splitter. The different terms primarily result from the methodology's use in different disciplines, such as ecology and economics, and is sometimes thought to correspond to how intelligent or rational modellers want to make the members of their virtual populations.

idealised picture that nevertheless isolates some important factors useful for answering the questions that I pose. If all goes to plan, the arguments of this thesis will demonstrate the utility of the view of scientific models presented in this section.

Call the view of scientific models I adopt the *simple similarity* view. This view follows in the tradition of Weisberg's (2013) *interpreted structures* view, itself a member of a family of similarity-based views including those of Ronald Giere (1988) and Peter Godfrey-Smith (2006). I distinguish my simple similarity view from Weisberg's as there are parts of his taxonomy that are omitted from mine. As the name suggests, this is a simpler similarity view of models than Weisberg's. These omissions and simplifications should leave me with a framework that is tailored to the project of this thesis by setting aside some of Weisberg's distinctions for which I have no use in the arguments to come. Furthermore, these omissions leave a framework that avoids some of the criticisms that have been directed towards Weisberg's view. Some of these criticisms are described in section 1.3.5

The core of Weisberg's view, which was proposed by Giere (1988), sees modelling in terms of a pair of relationships—specification and resemblance—between three entities: a model description, a model structure, and a target system. This basic framework is shown diagrammatically in figure 1.6.

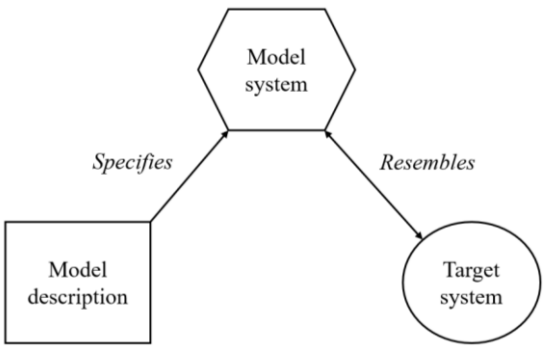

Figure 1.6 Adapted from Giere (1988).

In the following subsections, I will explain the parts of the simple similarity view, including model structures, model descriptions, the similarity relation, and model construals.

1.3.1 Model structures

First and foremost, a model must include a *model structure*. This can either be a physical, concrete object like a model organism or a scale replica of an Airbus A380, or it can be an abstract, mathematical object—a set of mathematical relationships—like a trajectory through multi-dimensional state space, a graph, and so on. The EBM, ESM, and virtual Physarum are all models with mathematical structures of varying complexity. A model structure is crucial because a modeller must have a thing that they can study and from which they can make inferences about their ideas or the world. As Cailin O'Connor and James Weatherall (2016) have argued, "structure" is an incredibly broad term, which you might as well replace with "thing." For the most part, I completely agree with this. Little harm comes from thinking of models as things—physical things or mathematical things—which are interpreted as representations of other things, including hypothetical things or classes of things. Yes, the sofa in the living room could be a model. In many cases, however, it is simply not a very good one (exceptional cases might include those in which the sofa is being used as a model of another sofa). The power of models lies not in what they *are* but what you *do* with them. Any old *thing* can be a model, but many things are not very good models for many purposes.

As I see it, the distinction between these two kinds of structure—physical and abstract—should be exhaustive. However, there are hybrid cases. For example, if a model consists of a physical robot in a physical environment and the robot's behaviour is coordinated by a computer program representing a possible cognitive architecture, then this is a hybrid case. In this hypothetical scenario, one part of the structure being investigated is the abstract computational relations driving the robot and another part of the structure is the robot's physical body and environment. Although hybrid cases are interesting, the analysis of computer simulations in this thesis will rarely if ever touch on concrete models, even hybrid ones. This is certainly *not* because few interesting comparisons can be made between concrete models and models investigated through computer simulation. Indeed, a not-insignificant part of the philosophical literature on computer simulation has concerned itself with comparisons between computer simulation models and scientific methods involving concrete objects (e.g. M. S. Morgan, 2005; M. S. Morgan & Radder, 2003; W. S. Parker, 2009). Rather, these questions

are left aside simply because I lack the time and space to speak satisfactorily to those comparisons. For the remainder of this thesis, I will be primarily if not completely concerned with scientific models with abstract model structures, like the three exemplar models described in section 1.2.

1.3.2 Model descriptions

The second part of the simple similarity view is a *model description*, which is used to specify a model structure. In the case of concrete model structures, model descriptions can take the form of blueprints, sketches, photographs, verbal descriptions, and more, and they play an important role in recording models and the results of model studies, communicating those results to others, and reconstructing the physical model. In the case of models with abstract structures, however, model descriptions play a special and essential epistemic role. While a model description can fade into the background when investigating a concrete structure, model descriptions move front and centre when investigating abstract structures. This is because, typically, the only way to rigorously investigate abstract structures is via the formal descriptions that specify them.

The illustrative cases from section 1.2 all have relatively different descriptions. The EBM is specified by the equations presented in section 1.2.1. The ESM and the virtual slime, on the other hand, are specified by lines of code in a programming language, like FORTRAN in the case of many ESMs and Processing in the case of Jones' slime mould. FORTRAN has long been used by scientists to code computer simulations, but Processing is a language designed to be accessible for use in the visual arts. Jones elects to use Processing partly because it allows for quick model construction, but also because of its visual capabilities, which allow him to easily construct slime-like figures like that shown in figure 1.4.

A reason to keep model structures and model descriptions separate is that modellers seem to be able to describe the same model structure with different descriptions. For example, the mathematical relationships of the EBM, which are specified with equations in section 1.2.1, can also be specified with lines of code. Alternatively, a computer simulation model can be described with

code from different programming languages. Indeed, this thesis will argue that much of the epistemic difficulties that emerge in simulation science can be understood in terms of uncertainties surrounding model descriptions and the structures they are intended to specify. In Chapter 2, for example, we will see that numerical approximation of continuous mathematical equations has long been a source of uncertainty in climate modelling. This serious difficulty can be understood as an epistemic problem that arises when modellers are forced to change a model description while still attempting to specify that same, or a very similar, mathematical structure. Likewise, in Chapter 5, we see that agent-based modellers face similar difficulties discerning what abstract structures are actually being investigated by their colleagues on the basis of the information provided in publications, which hampers efforts to replicate model studies, an important method for ensuring the quality of model results.

Some philosophers (e.g. Odenbaugh, 2015, 2018b) may wish to dispense with the distinction between model structure and model description, claiming that the model just is the equations or lines of code. Again, my project is not to produce a deep metaphysics of models (a view shared by Weisberg, 2015). Rather, I make use of the distinction between model structure and model description because it is a useful piece of conceptual apparatus in a discussion of the epistemology of modelling with computer simulation. Perhaps these observations can all be expressed without the distinction between structures and their descriptions, but they are easily expressed with it. More details about model structures and descriptions will be left for chapter 5, which, to a large extent, focuses on this topic.

1.3.3 The similarity relation

The next relationship in the simple similarity view is the view's namesake: the similarity between model structure and target. At least initially, it might seem that no abstract structures will be similar to their representational targets because these targets—targets like the climate or the slime mould—are physical things (Hughes, 1997; Odenbaugh, 2015, 2018b). How on Earth is a computer program anything like a slime mould? Indeed, the worry about similarity between mathematical relationships

and physical systems seems to have been a key motivation for Godfrey-Smith (2006, 2009b) to endorse a models-as-fictions view of mathematical models, according to which the model structures of mathematical models are not sets of mathematical relationships, but imagined, fictional worlds—think Middle Earth, Hogwarts, or Westeros—which would be concrete if only they were real. Hard to see how a computer program resembles a slime mould, easier to see how an imagined slime mould resembles a real one.

I do not think we need such a view to make sense of similarity.[8] Instead, we can recognise that scientists do not typically compare their mathematical models directly to the world. Rather, they compare the model to a mathematical representation of the target or to the mathematical relationships embodied by the target. Consider the images of the virtual and biological Physarum in section 1.2.3. Yes, the pictures look similar, but this similarity can be understood in terms of the mathematical relationships of the graph: the number of nodes in the network, number of edges between nodes, distance between nodes and length of edges. Yes, they are both green, but this carries no epistemic weight. Likewise, if we want to know how similar an ESM is to the actual climate, we compare the data it produces to actual and historical climate data. Hence, like is compared with like—that is, model data is compared with empirical data and quantities are compared to quantities—and the mystery dissipates. To reiterate the view, models are not directly compared with the physical world but with representations of targets (Woodward, 1989, 2011): an ESM may be compared with data about average temperatures or precipitation, and Jones' virtual slime may be compared with data about real slime networks. This view of is shown diagrammatically in figure 1.7.

[8] However, I do think that imagined worlds or scenarios play an important role in model construction and can assist a modeller in determining what kinds of relationships ought to be included in a model.

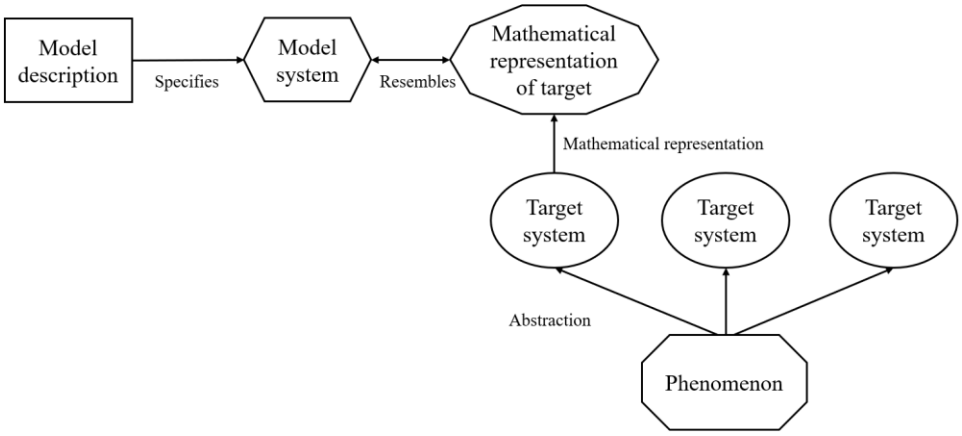

Figure 1.7 Weisberg's view of the relationships between model systems and their targets via mathematical representations of targets (adapted from Weisberg, 2013, p. 96).

As that diagram shows, there are two distinctions to be made. The first is between the raw phenomenon and target systems. Target systems can be understood in slightly different ways, with Weisberg describing them as abstractions of raw phenomena. That is, they are the isolated parts and properties of a phenomena that are of interest to the scientist given the context of their investigation. Alkistis Elliott-Graves (2020b) has a somewhat similar view, building on accounts like Weisberg's, but sees target systems as "those aspects of the real-world system that are studied in order to gain knowledge about the phenomenon" (p. 8), dropping the explicit reference to abstraction. For example, the raw phenomenon of slime mould behaviour might well involve molecular activity. It almost certainly does. However, these molecular parts and properties are not of interest to Jones and so do not factor into his target system. This is not just a matter of scale either. Features of the slime mould, like its complete developmental and reproductive cycle, are not of interest to Jones and so are not included in his target system.

Elliott-Graves (2020b) has developed an account of specifying a target system, which involves: (1) fixing on a domain of study comprising the spatial and temporal location of interest

within the raw phenomenon; (2) partitioning the domain of study into parts and properties; (3) determining which parts and properties are relevant given the modellers' purpose; and (4) omitting parts and properties that they deem irrelevant given their purpose. At this point, they have specified a target system, which may be more or less apt for their goal. It's entirely possible that they have omitted a part or property that they ought not to have, and Elliott-Graves' account also deals with this aspect of target system evaluation. However, let me now move on to the second distinction.

The second distinction shown in figure 1.7 is between target systems—all and only those parts and properties of the real-world phenomena deemed to be relevant for the study—and models of data or mathematical representations of the target system. These are most often the measurements taken of the parts and properties isolated in the target system. For example, these might be data sets describing the graphical properties of a slime mould's nutrient network, such as path lengths and number of nodes. It is to these representations of targets systems to which model results are compared and to which evaluations of similarity are applied.

Does the simple similarity view endorse Weisberg's (2013, Chapter 8) set-theoretic feature weighting account of similarity? Once again, this is an aspect of Weisberg's view that I will set aside. To briefly describe this piece of conceptual machinery, Weisberg presents a measurement of similarity between model and target which decomposes both into features, tallies the features they share and those they do not, and adjusts the weight of these based on which features are deemed important. This is based on Amos Tversky's (1977) contrast account of similarity. To fine tune the account for the model-world relationship, Weisberg introduces a distinction between two kinds of features. First, attributes are the system's properties or states, such as the intensity of the albedo factor or atmospheric transmissivity in the EBM. Second, transition rules are the functions or procedures that govern how the attributes are updated, like the relationship between the albedo factor or atmospheric transmissivity and global average temperatures in the EBM, which are intended to be read causally. With all the features of the model and target collected and weighted, we can use Weisberg's set-theoretic equation to determine the similarity between model and target.

The view itself has seen some criticism from multiple angles. Wendy Parker (2015) presents a thorough response to Weisberg's machinery, and some of her critiques are picked up and expanded upon by Wesley Fang (2017). Fang makes two arguments. The first I interpret as a kind of argument from practice: scientists do not need an abstract set-theoretic measurement of similarity because they have their own statistical methods, such as maximum likelihood estimation, with which to determine the fit between model and target. The second argument follows naturally: the methods that scientists use to determine fit between model and target are holistic while Weisberg's account is atomistic, so it does a poor job of describing how scientists approach similarity even in the abstract. I would find the first argument compelling if it were clear that Weisberg's only intention was to provide scientists with a formalism for measuring similarity rather than, say, providing philosophers of science with a formalism for measuring similarity that they could then use—and here is the important part—for drawing the distinctions between kinds of models that philosophers of science find interesting. For example, Weisberg uses his set-theoretic feature weighting account to distinguish between "hyperaccurate modelling," "how-possibly modelling," "minimal modelling" and more (2013, pp. 150-151). Without going into the details of these categories, Weisberg clearly intends for his account to benefit philosophers with their taxonomic analysis, and this value would remain irrespective of whether practicing modellers used other statistical techniques for making their similarity judgements.

The second argument I find less convincing simply because modellers often do compare specific parts of their model to the target even if some of the statistical methods they use are not sensitive to these decompositions. For example, climate modellers might claim that their models do a good job of reproducing global average temperatures from a historical data set but that they are not reproducing precipitation patterns near the equator correctly; some attributes of the model are similar to attributes in the target while others are not. Likewise, they might argue that part of the problem is that their cloud parameterisation schemes need improving; some of the model's transition rules do not match the causal processes in the target. Moreover, to make a convincing argument that Weisberg's atomistic feature weighting account is misguided because practicing modellers use holistic similarity measurements, Fang would need to show that all or even most of the statistical (and non-statistical)

methods used by scientists are holistic. While Fang certainly makes the convincing point that similarity judgements are sometimes holistic—where maximum likelihood estimation is used—we would need to see a more exhaustive exploration of scientific methods to justify the claim that Weisberg's atomistic theory does not match the holistic practice of modellers. Still, while I think it is useful to talk about which parts of a model are similar to a target and which are not, I think Weisberg's formal machinery for measuring similarity is less so and I will not be using it in this thesis.

With the key parts of the simple similarity view introduced, I will now describe the model construal or model interpretation, which turns a mere structure—just a thing—into a model.

1.3.4 Model construal

A model construal is responsible for turning things into models. The first part of a model construal or interpretation is the *model assignment*. The model assignment tells us what parts of the model are supposed to represent: α in the EBM represents the albedo factor; the ESM's dynamical core represents the atmosphere and ocean; and Jones' digital agents represent hypothetical units of gel/sol interaction, collectively representing Physarum's body and the flow of Sol throughout.

Not all parts of a model need to be assigned or interpreted as playing a representational role. For example, when the virtual slime is in its oscillatory condition, each agent keeps count of the number of times they have tried to move forward but have been blocked by other agents. The agents use this count to try moving into an empty space increasingly far away. Although the behaviour or pattern this procedure produces represents the accumulation of force within the contracting and relaxing Physarum, the procedure itself is not mapped onto anything in the target—there is no count being taken in the real slime mould. Instead, the dynamics that result are intended to correspond to the world since the procedure produces behaviour resembling fluids acting under pressure. Likewise, the coupler in an ESM does not represent any aspect of the target but instead serves the function of ensuring that the other components, such as the ocean and atmosphere components, which do serve a representational role, are able to interact and represent the combined ocean-atmosphere system.

The second part of a construal is an *intended scope*. This specifies the range of targets for which the model is supposed to be a representation. In terms of the distinctions introduced in section 1.3.3, the intended scope specifies the phenomenon, the target system consisting of relevant parts and properties of that phenomenon, and which of those relevant parts and properties the modeller intends to represent with their model. For example, the virtual plasmodium's intended scope is one particular system: the slime mould *Phyrasum polycephalum*. Moreover, it is also only supposed to represent that system in one part of its lifecycle—the plasmodium stage, in which the ameboid eats and grows—and so does not aim to represent anything about, for example, how the slime mould reproduces. The model is also supposed to represent the system at a particular level of description: the model represents the actomyosin matrix of which the slime is composed because Jones intends to investigate how collective behaviour at this level of organisation can lead to the formation of structures at a higher level of organisation—that of nodes and edges in a nutrient network. Jones does not intend to represent anything below the material level within his model, so the processes that underly the behaviour of entities at the material level can be black-boxed and the behaviour of the hypothetical Gel/Sol units merely described.

Since the lower level of organisation is outside Jones' intended scope, it is not a strike against Jones' model that it misrepresents or fails to represent anything at these lower levels. That is, unless those mis- or non-representations create problems at the higher levels of organisation that concern Jones. Consider, for a moment, our climate models. ESMs represent a host of underlying physical and climate processes omitted from simpler climate models like the EBM. One might worry that omitting these extra processes leads to inaccurate results if these processes interact in ways that alter the behaviour of the larger climate system, in which case a more comprehensive model can be used to test this possibility. If this testing reveals that the simpler EBM remains accurate despite the omissions, then the EBM can continue to be used within its original intended scope. If it is shown to be inaccurate, however, the model may either be thrown out entirely or, as is more common, the intended scope typically adopted when using the model will change, restricting the scope of application to only those cases where it is accurate.

While models like ESMs and Jones' virtual Physarum are intended to represent particular systems over particular temporal ranges and at particular levels of organisation, models often have a broader scope and are intended to represent general classes of targets. The 0D EBM, for example, need not represent one target, like the Earth, but planets more generally: Earth, Mars, Venus, and so on. When the EBM is instantiated with empirically determined values, then it will be intended to represent the more specific target from which the empirically determined values derive. Similarly, if we were to use Jones' virtual plasmodium for the purposes of Physarum-based computing rather than Physarum-based modelling—that is, to use slime-like behaviour to solve computing problems rather than to use representations of the slime to better understand the slime—then it would have a broader scope than it would have otherwise.[9] In this case, the model might be intended to represent or apply to any case where there is, for example, a travelling salesman problem to be solved, such as an Amazon warehouse or delivery service.

The final part of a model construal is the *fidelity criteria*, which state the desired resemblance between a model and target given a modeler's purpose and the model's intended scope. There are two kinds of similarity for which a modeller may aim with their model: dynamical and structural. Dynamical similarity concerns the match between states of a system changing over time, or the output of a system given some input. Structural similarity, on the other hand, concerns how well the model represents the target's underlying causal structure. To demonstrate this distinction, let's go through an example.

[9] One might think that in such cases, the virtual slime would be more akin a calculation device, like an abacus, rather than a model. I am sympathetic to this view, though it is worth noting that, for any given problem that the virtual slime could be used to solve, such as determining efficient pathways through a warehouse or through a transportation network, the environment in which the Physarum is placed would have to be a model of the target problem space. Still, at this point we might not think of the virtual slime as serving any representational purpose but simply serving the function of calculating a solution within that problem space.

ESMs are intended to dynamically resemble the Earth system. If we introduce a forcing scenario to an ESM—perhaps we double atmospheric CO_2 relative to preindustrial levels—we want the model's response to match the response of the real world. So, if doubling CO_2 increases global temperatures in our model by, say, about 2 degrees Celsius, then the model is dynamically similar to the target if the Earth's temperature would also increase by about 2 degrees in response to CO_2 doubling. Dynamical similarity is intuitive enough, but what does it mean to say that a model is similar in terms of a target's underlying causal structure? All this means is that the relationships in the model responsible for determining its behaviour have counterparts in the target. For example, in the EBM, treating the planet as a black body is a sufficiently close approximation, but the planet is not a black body and does not emit infrared radiation at the same rate.

1.3.5 Complications

Before moving to the details of how exactly I am understanding the term "computer simulation," let me address two worries that one might have with my adoption of the simple similarity view of scientific models.

First, recall that I have adopted the distinction between models with abstract structures and concrete structures, focusing on abstract structures in this thesis. Weisberg makes a further distinction between two kinds of abstract structure: mathematical and computational. I have opted to leave this distinction aside. This is because I find that "computational model" in Wiesberg's sense can confuse what I see to be the real points of interest in a comparison between abstract model structures that are investigated with computer simulation and those that are not. Instead, if I ever use the term "computational model," I simply mean an abstract model structure that is being investigated through computer simulation, though I will often use the more explicit "computer simulation model." Again, the details of my understanding of these terms will be discussed in section 1.4.

Now, why not adopt Weisberg's distinction? From the names alone, such a distinction might appear quite useful for my project: does Weisberg's distinction between mathematical and

computational structures hold a key to the epistemology of computer simulation? First, let me quickly summarise the distinction. On Weisberg's view, computational structures are algorithms or procedures, which are a proper subset of mathematical structures. This means that computational models are also a special subset of mathematical models; however, they are mathematical models where the algorithm provides the explanatory power as opposed to, say, the relationships between variables or parameters. On such a view, it is possible to create models that are investigated through computer simulation but are not computational models in Weisberg's sense because the algorithm plays no special explanatory role. ESMs are like this. If you wanted to explain, say, the circulation of the atmosphere, the equations that are numerically approximated in the model do the explanatory work, not the algorithm.

One reason for leaving the distinction is that it has been criticised as something of a distinction without a difference (O'Connor & Weatherall, 2016). For example, O'Connor and Weatherall argue that models can sometimes be described with both continuous mathematics and in discrete, algorithmic form, without much of interest changing. For example, while the 0D EBM as it is presented in section 1.2 is not algorithmic, we can create a computer program that encodes those mathematical relationships. Since I'm dropping the distinction, I will not spend much time defending it, but I will note that I think O'Connor and Weartherall's criticism misses the mark here. In their discussion, they argue that the same or very similar mathematical relationships can be specified with equations or algorithms. In doing so, however, they have not considered whether the procedure plays any important explanatory role. If it turns out that a 0D EBM or the replicator dynamic (their example) can be expressed algorithmically, but there is little explanatory gain from thinking of it in this form, then there is no reason to think of it as a computational model.

Second, one might have a more general worry about the adoption of a similarity-based view of modelling. Wendy Parker (2020b), for example, has argued that models should be judged not in terms of their similarity to a target or targets, but in terms of adequacy-for-purpose. This might be thought of as a specific instance of a broader *tool-based* or *use-based* view of modelling (e.g. Currie, 2017; M. S. Morgan & Morrison, 1999), which sees models not, first and foremost, as representations that are

more or less accurate or similar, but as tools that can be more or less apt for a task. Although an alternative view of models in the literature, Parker's account is not incompatible with the simple similarity view as I have presented it. Parker's account is motivated by the idea that models should not be evaluated solely in terms of similarity or any other view of representational accuracy, arguing that representational fidelity is only one part of adequacy-for-purpose. Additionally, models must stand in the right kinds of relationships with the user, the methodology, the circumstances, and the modelling goal (W. S. Parker, 2020b, p. 464). In many cases, these other relationships have an impact on the desired degree of similarity. For example, a lower fidelity representation might be preferable to a higher fidelity one if the higher fidelity model is too computationally demanding to sufficiently explore the range of scenarios desired by the modeller (p. 466). In such cases, these factors can be incorporated into the simple similarity view via the construal, such as in the form of the fidelity criteria.

Perhaps more problematically for a similarity-based view of models is that there appear to some cases—in fact, many cases—in which the model's purpose is to assist a modeller with thinking through their ideas or assumptions rather than to predict, manage, explain, and so on, the behaviour of target. In such an instance, the model appears to be directed toward ideas or hypotheses rather than a target in the world. Consider, for a moment, John Maynard Smith's (1964) haystack model of altruistic behaviour by group selection. This model is built to demonstrate that such behaviour can only evolve in very specific situations, where populations are frequently divided into small and completely isolated groups in which the altruistic traits can gain a foothold where they would otherwise quickly disappear in a mixed population. In this instance, Maynard Smith's model appears to be directed either toward some hypothetical target of a population of mice that live only in a set of haystacks which are periodically isolated and then brought together allowing groups to mix, or toward hypotheses and assumptions about reproduction and population genetics. Such a case may appear problematic for a similarity-based view of modelling as the evaluation of adequacy-for-purpose appears not to be describable in terms of similarity to a target but in terms of exemplification of assumptions.

As Elliott-Graves (2020b) argues, even in cases like these, there are target systems that enter into the evaluation of the model. This is because, once the model has been constructed and its results analysed, inferences are made back to the world (see also MacPherson & Gras, 2016). For example, Maynard Smith concludes his paper by stating that "Whether this is regarded as an argument for or against the evolution of altruistic behaviour by group selection will depend on a judgment about how often the necessary conditions are likely to be satisfied" (p. 1147). To put it another way: the model and the conditions it instantiates can be applied to any relevant target (i.e. reproducing populations) to determine whether altruism is likely to have evolved by group selection. In many cases, Elliott-Graves argues, models appearing to be directed toward ideas or hypothetical scenarios still have very general targets, such as evolving populations in general, as in the case of the Haystack model.

For now, I will focus on the model-target relationship in the view of model-based science I have presented in this chapter on the grounds that this relationship is central to three of the five substantive chapters ahead. Chapter 2 examines the capacity of computer simulation to permit the investigation of models with greater representational fidelity than would otherwise be possible. Chapter 4 similarly looks at the capacity for computer simulation to allow the representation of target structures that could not otherwise be represented. And Chapter 6 looks at the constraints placed on high-fidelity modelling when modellers do not share one or even a small set of targets. The remaining chapters—Chapters 3 and 5—focus on model structures and model descriptions, the other key aspects of the simple similarity view. Hence, while I endorse the lessons of Parker's adequacy-for-purpose view and take my account to be completely compatible with such a view, I will continue to use the simple similarity view as it will assist me with my particular philosophical project. As you will see, this thesis does focus, to a large extent, on descriptive and predictive models where, so the simple similarity view suits this purpose. The reason I focus on descriptive and predictive models here is because this is one area of modelling where computer simulation models can make their greatest and most interesting contributions.

1.4 What is a computer simulation?

So far, I have presented some exemplar models, two of which were computer simulation models, and a framework for conceiving of models as structures specified by descriptions and resembling targets. But this thesis is not only about the epistemology of models in general, but about the epistemology of computer simulations in particular. Before concluding this chapter, then, I ought to say something about what I mean by "computer simulation" and "computer simulation model."

Within the philosophy of science, there has been a great deal of disagreement about the defining characteristics of computer simulation. I'll present some of these suggestions here before stating my own, comparatively deflationary, position. One suggestion was that computer simulation is characterised by the numerical approximation of analytically intractable equations using digital computers (Humphreys, 2004, p. 49; Eric Winsberg, 1999, p. 275, 2001, p. 444, 2003, pp. 107–108, c.f. Frigg and Reiss, 2009, p. 596). Although many simulations fit this description—think of the dynamical core of an ESM—many simulations do not. Agent-based models like Jones' virtual slime are counterexamples to this trend. In these cases, computer simulation can be used to perform calculations that would simply take too long to complete by hand. Although the point about numerical approximation of continuous mathematics is off the mark, there is an important point to note here: simulation is often used to manipulate abstract structures that are too complicated or complex to be manipulated with pen, paper, and brains alone. This fact—that simulations deal with big model structures—has interesting consequences for their epistemology that will explored in this thesis, especially in Chapters 2 and 6.

Another suggestion was that an essential aspect of computer simulation is the fact that they are temporal. Parker (2013a) defines a computer simulation as a time-ordered series of states that represents another time-ordered series of states. For example, the evolution of the virtual slime or virtual atmospheric circulation represent the evolution of Physarum as it forms its nutrient networks or the actual circulating atmosphere. Unfortunately, this definition excludes instances where computers are used to investigate static models. For example, statistical models, which often involve sampling from probability distributions to produce, say, phylogenetic trees, do not represent the dynamics of

their targets. A phylogenetic tree is intended to represent the structure of its target—that is, the relatedness of different species—rather than dynamics. Although following the algorithm and making the requisite calculations takes time and happens in time, the corresponding time-ordered states do not represent time-ordered states of the target in the same way that the changing states of a climate model represent the changing states of the corresponding parts in the target.[10]

Paul Humphreys (2004) has responded to this line of criticism by arguing that even the computational investigation of static models is temporal as the calculations performed by the computer take place in time even if they do not serve a representational function in doing so. It's hard to see why this is particularly unique to computer simulation, however. Long before the advent of the digital computer, mathematicians performed their calculations in time rather than instantly. Humphreys is also right that time does create constraints as we do not wish our computer to perform so many calculations that it will take an inordinate amount of time and render the results useless. Consequently, simplifications, approximations, and idealisations may be added to models to ensure that their computational manipulation can be completed quicker. But, once again, this is not especially unique to computer simulation as considerations of tractability have long motivated the addition of simplifications, approximations, and idealisations into models.

Another suggestion has been that visualisation is a defining feature of computer simulation. Visualisations, here, are understood as movie-like animations that can accompany the calculations underlying the simulation, showing the evolution of the virtual system as it changes. These animations might show how Jones' particles slowly form their network structures or show the circulation of the atmosphere and oceans in an ESM. Humphreys argues that the use of visualisation is representationally distinctive of simulations (Humphreys, 2004, pp. 112–114), while R.I.G. Hughes

[10] Note that phylogenetic analyses can also specify the order of trait changes, gains, and losses on the tree. These appear to be facts about dynamics, but they are still represented statically. Contrast this with an Earth system model where changing state variables are updated at every time step and play a representation role at every time step.

(1999) distinguishes between genuine computer simulation and mere numerical calculation on the basis that visualisation techniques give simulations a mimetic quality. This seems to me to greatly overstate the importance of visualisation. Instead, I agree with Eric Winsberg's view that visualisation techniques are just one technique among many used in the data analysis process to make the stack of numbers that computer simulations produce more intelligible to human scientists. For many of the purposes to which simulations are put, however, the visualisations are inessential. For example, if we want to use ESMs to project possible future average global temperatures, then what we want is a graph of the change in this value over time rather than a movie of fluid dynamics. Not only does visualisation appear inessential for simulation, it is also does not appear to be unique to simulation. Mathematical modellers often draw diagrams of simple curves and lines to help them get a handle on their abstract objects (Frigg & Reiss, 2009, p. 607; M. S. Morgan, 2012). Sometimes pictures help us get a grip on abstract objects. Simulation is no different, but the pictures might be proportionally more complicated relative to the abstract structures being studied.

For as long as modellers have used computer simulations, they have referred to their simulations as computer experiments. British meteorologist Eric Eady, for example, reacted to Norman Phillips' 30-day forecast, one of the first of its kind in 1955, by stating that "Numerical integration of this kind… give[s] us [the] unique opportunity to study large-scale meteorology as an experimental science" (c.f. Heymann & Achermann, 2018, p. 613). Simulation undeniably has some similarities to experimentation. Most obviously, one needs to run a simulation many times—that is, one needs a computer to execute the model's algorithm many times—to fully explore the consequences of the algorithm. This is because these algorithms frequently involve stochasticity, leading to different results on different runs and, even when the algorithm is completely deterministic, different initial conditions can produce varied and interesting results. Moreover, modellers typically want to know how their virtual systems will behave given a range of different parameter settings, so will need to run many simulations to explore the consequences of these altered settings. Further similarities are the techniques, like visualisation, that must be used to analyse the large data sets that

all these simulations produce, and the tinkering with physical instrumentation and the associated error management (Dowling, 1999; Winsberg, 2010).

At its most extreme, these similarities with experimentation have led some to view simulation as "an entirely new mode of scientific activity," claiming that the "methodology lies somewhere intermediate between traditional theoretical physical science, and its empirical methods of experimentation and observation" (Rohrlich, 1990, p. 507).[11] For example, modeller Robert Axelrod (1997, pp. 3–4) described agent-based modelling as a "third way" of doing science between traditional inductive and deductive approaches. More abstrusely, Sergio Sismondo has claimed that computer simulations uniquely "occupy an uneasy space… between abstract and concrete…" (c.f. Frigg & Reiss, 2009, p. 607; Sismondo, 1999, p. 247).

In comparison to these claims, my own view of simulation is rather deflationary. Like Roman Frigg (2010), I understand computer simulation as a method for investigating models described algorithmically. An individual simulation or "run" occurs when the model's algorithm is executed. A "simulation study" refers to many runs and their output. When I speak of a "computer simulation model," I mean the program being investigated. I will often use "computer simulation" loosely to refer to the programs and their analysis unless it is imperative that I specify only one part of this process. Given that computer simulation has been given such a deflationary definition here—as a model with a description that can be manipulated by a digital computer—it might appear that there is not a whole lot that differentiates modelling with computer simulation to modelling with pen and paper. Indeed, I take whatever we say about the epistemology of computer simulations in particular to tell us something about the epistemology of modelling more generally because computer simulations are just a genus of model. However, even this deflationary definition of computer simulation directs our attention to a set of features which create interesting and important epistemic challenges. Below, I will conclude this chapter by stating some of these challenges, directing you to their location within the thesis.

[11] Others holding this view include simulation modeller Norman Zabusky (1987), sociologist of science Deb Dowling (1999) and historian Peter Galison (1996) (c.f. Winsberg, 2010).

1.5 Conclusion

In this chapter, my goal has been to introduce three different kinds of scientific model: a simple mathematical model, a class of highly complex computer simulation models, and a far simpler, though still complex, computer simulation model. These case studies, along with others to be introduced in future chapters, will assist me in answering my primary research questions about the impact digital computers have had on model-based science. In addition to these exemplars, I described a framework for talking about scientific models. According to this view, modelling is understood as involving: (1) model structures, which are comprised of mathematical objects in the case of mathematical and computer simulation models, or concrete objects otherwise; (2) model descriptions, which are the equations or lines of code used to specify the relationships of the model structure, or pictures, blue prints, and more in the case of concrete models; (3) a target system and representation of the target system to which the model is compared; and (4) a construal, which is used to determine how similar a model ought to be to a target relative to the modeller's purpose.

In the coming chapters, I will use this framework in the following ways. In Chapter 2, I explore the consequences of building models with very complex and complicated model structures, which could feasibly be investigated only with computer simulation, with Chapter 3 continuing this analysis and examining the method of using multiple models with similar structures to reduce the opacity of models. In Chapter 4, I will examine the ability of models with particular kinds of structures—that is, agent-based models—to represent targets at a level of organisation which would not be possible without computer simulation. Once again, these could not be investigated without computer simulation. In Chapter 5, I consider the problems that can arise from the ability to describe simulation models with many different programming languages and the opacity this creates within the modelling community. Chapter 6 returns to the topic of highly complex and complicated models aimed at predicting and managing natural and social systems, examining the conditions, including community infrastructure, that must be in place to undertake these costly endeavours successfully. A key conclusion is that cases where many modellers share a target or set of similar targets have a better

chance at succeeding because the ability to concentrate resources supports the kinds of epistemic activities described in Chapters 3 and 5.

2

The strategy of model building in climate science I: The trade-off between comprehensiveness and comprehensibility in climate science

This chapter makes a claim about the epistemology of computer simulation: computer simulation enables the investigation of abstract model structures that are more complex than ever before, but this complexity comes with its own set of epistemic challenges. This argument uses Richard Levins' seminal work responding to the emergence of simulation modelling in ecology in the 1960s and applies it to high-fidelity climate models, demonstrating how his insights remain relevant today.

2.1 Introduction[12]

One obvious way in which computer simulation has changed model-based science is that it permits the investigation of abstract structures that are more complicated and complex than ever before.[13] But is this always a good thing? This chapter will argue that, although these new abstract structures can represent systems in greater detail, their complexity comes with its own set of epistemic challenges, and these prohibit them from ever replacing simpler mathematical representations.

To make this argument, this chapter will use theoretical biologist Richard Levins' (1966) paper "The Strategy of Model Building in Population Biology" as a scaffold. Despite its age, its

[12] Some of the material included in this chapter has been published in (Walmsley, 2020).

[13] A note on complicatedness and complexity. To be precise, a more complicated system is one with more parts and subparts and subsystems operating together within the system. Complexity enters the picture as the interactions between these parts and subsystems become non-linear and more difficult to predict. However, I will use "complexity" as shorthand for the combination of both unless otherwise stated.

lessons remain incredibly relevant to model-based and simulation-based science today. In the 60s, Levins argued that it would be naïve to think that we—living in an imperfect world and without god-like powers—could ever create perfectly realistic representations of our world. And yet it appeared to Levins that this "brute force approach" was precisely the philosophy behind a then-new research paradigm in biology. Systems ecologists like Kenneth Watt argued that the simple mathematical models Levins and his colleagues built were too unrealistic to reveal anything of worth about real-world systems (Watt, 1956). Watt claimed that only computer simulations could capture all the relevant features of ecological phenomena, so we should trust them more than simple mathematical models. For Levins, on the other hand, the brute force approach could not be pursued unaccompanied by simpler models in virtue of three epistemic challenges that highly complex models face: they have large appetites for data that may be difficult or impossible to acquire; they are computationally intractable; and they are incomprehensible to the human modeller. In this chapter, I'll argue that the brute force approach appears to be commonplace in climate science and that the problems that Levins described still apply. [14]

As I will argue, the three epistemic challenges Levins described are still faced in state-of-the-art climate modelling: (1) historical data sets have gaps and the procedure used to fill the gaps, reanalysis, is an additional source of uncertainty; (2) computers remain insufficiently powerful to resolve important but poorly understood sub-grid processes; and (3) the complexity of climate models obscures their internal causal structure, making them difficult to evaluate. I will argue that this creates a modelling trade-off, which must be addressed with multiple models of varying degrees of complexity, and not just ensembles of highly complex climate models.

In making my argument, I will be contributing to the philosophy of science and modelling literature that has formed around Levins' work. Within this literature, John Matthewson (2011) has

[14] Levins also argued that such models lacked generality. This is not a criticism I will look at in detail in this chapter because I will be arguing, in section 2.3, that high-fidelity climate models do not need a high degree of generality.

argued that modelling trade-offs are particularly problematic in Levins' own field of population biology due to the heterogeneity among ecological target systems: if you pick two ecosystems at random, like the Great Barrier Reef and Yosemite National Park, they are unlikely to share all their important causal features. Making a model more realistic or precise with respect to one target system, then, necessarily reduces how general it is by misrepresenting targets that do not share the properties described in the model and possessed by the first target (Matthewson & Weisberg, 2009). This view implies that modelling trade-offs will not be as severe in climate science because modellers frequently deal with the one target system over a relatively short period—Earth between 1850 and 2100—eliminating the problem of representing a set of causally heterogeneous targets.[15]

Instead, I will argue that the trade-offs faced in climate science are better explained according to Jay Odenbaugh's (2003, 2006) view that the three epistemic challenges facing brute force models create a demand for simpler models that require less data, require less powerful computers, and, most importantly, are easier to understand.[16] Consequently, the trade-off in climate science is primarily between realistic and precise models with predictive power and much simpler and highly idealised models that facilitate understanding (Charney, 1963).

[15] You may worry that time is enough to generate causal heterogeneity. I will be addressing this possibility in the arguments presented in section 2.3.

[16] Comprehensibility is not always independently desirable. If the purpose of the model is simply to forecast and assist in system management, then comprehensibility is desirable to the extent that it assists modellers with these goals. It may, for example, assist the modellers with debugging their simulation code and ensuring that it is behaving as expected. However, understanding and gaining knowledge *about the world* is its own scientific end, which is better served by models oriented toward this purpose. I will say more about models that are themselves comprehensible and facilitate understanding of worldly processes in Chapter 3. Here, I will just note that, if scientists understand a phenomenon better, it will help them with their modelling. This might take place in the form of, for example, knowing what parts and properties need to be captured in a target system and model and which can be omitted given the scientists' goals.

Here is the plan for the chapter. In section 2.2, I will describe Levins' desiderata, which he believed could not be jointly maximised and trade off against one another. In section 2.3, I present Matthewson's argument that the trade-offs are generated through target heterogeneity rather than target or model complexity. I claim that, if he is right, we should not expect to find the trade-offs in high-fidelity climate modelling because these models share a target. In section 2.4, I demonstrate that high-fidelity climate models, like the Earth system models (ESMs) described in chapter 1, are an instance of brute force modelling. In section 2.5, I argue that their complexity creates a trade-off that has long been acknowledged within the climate modelling community.

2.2 Modelling trade-offs and heterogeneity

The primary message of Levins' paper was the defence of a kind of model pluralism. Levins argued that no one model would be suited to perform every role that we might like our scientific models to perform. Just as a hammer and screwdriver perform different roles but each earn their place in the toolkit, so it is with models. For Levins, there were at least three desiderata a modeller might wish to satisfy with their models depending on their different aims (1966, p. 422):

> It is of course desirable to work with manageable models which maximise generality, realism, and precision toward the overlapping but not identical goals of understanding, predicting, and modifying nature. But this cannot be done.

On this view, we cannot produce a single model that maximises generality, realism, and precision because these desiderata compete such that modellers can only maximise two at a time. This creates a trichotomy of models that each maximise a pair of desiderata. In this section, I will describe these desiderata in a little more detail and provide some further description of the trade-offs.

Levins did not provide a detailed description of these modelling desiderata, but such descriptions have been articulated by Weisberg (2004, 2006a). We can distinguish between two senses of "realism" (Weisberg, 2006a). On the one hand, a modeller could aim to build something that represents as much of the target's causal structure as possible. That is, as many of the causally relevant

processes in the target are represented as functions or algorithms in the model as possible. Alternatively, they could aim toward building a model that reproduces the target's dynamics as closely as possible, underlying causal structure be damned. Typically, however, a model aiming toward a high degree of realism aims toward representing both the relevant causal relationships and doing so such that the dynamics of the model match the dynamics of the target.

Weisberg describes precision as "the fineness of specification of the parameters, variables, and other parts of the model descriptions" (2006a, p. 636). To put it simply, saying the Earth will warm by 2.35 degrees Celsius by 2100 is more precise that saying it will warm by between 2 and 3 degrees Celsius by the same date. While a modeller may in many cases prefer precise results, there are some cases where imprecise results suffice or are preferred (Elliott-Graves, 2020a). False precision can be even worse than accurate imprecision. Would you rather think your taxes were due in late September (imprecise) or on October 10 (precise) if it turned out they were actually due by September 25? In some circumstances, such as in climate science, uncertainties are such that increasing the precision of the model often leads to false precision (W. S. Parker & Risbey, 2015). In cases like these, greater precision comes at the cost of accuracy, so a modeller may sacrifice some precision to preserve accuracy.

Finally, Weisberg distinguishes between two kinds of generality. A-*generality* refers to the number of actual target systems a model describes, and *p-generality* refers to the number of "logically possible" target systems a model describes. Weisberg (2006a, pp. 634–635) argues that, although Levins explicitly endorsed *a-generality*, he implicitly endorsed *p-generality*, using examples such as Wright and Fisher's one-locus models that represent possible but not actual biological systems. Getting a sense of the landscape of possibility can assist in explaining, for instance, "that a different history could have led to a different reality" or "why our world cannot have the model system and what laws of nature would have to change in order to make this merely a contingent fact" (Weisberg, 2013, pp. 128–129 see also pp. 121-129). Indeed, a modeller may wish to explore a model system that is biologically impossible precisely to discover the reasons for this impossibility. This would be the equivalent of a perpetual motion machine in physics. Similarly, a modeller may wish to show how

something is biologically possible but very unlikely, as John Maynard-Smith (1964) did with his haystack model and the evolution of altruism by group selection. The kind of generality a modeller aims toward—possible or actual—will depend upon their research goals.

According to Matthewson, the trade-offs between these desiderata, and their unique force within biology, are generated by the *heterogeneity* among different ecological systems. And, as Matthewson notes, Levins was aware of heterogeneity in his original paper: "the multiplicity of models is imposed by the […] demands of a complex, heterogeneous nature…" (Levins, 1966, p. 431; c.f. Matthewson, 2011, p. 326). We could build a maximally precise and realistic model of an Airbus A380 that describes the causal structure of all the A380s, Matthewson argues, because their causal structure, though complex, is homogeneous (Matthewson, 2011, p. 331). This means that, if you have modelled one, you have modelled them all, irrespective of the amount of detail included in the model. Having a complete replica of the Great Barrier Reef ecosystem, however, does not give you a complete replica of the Yellowstone National Park ecosystem because the two complex systems are causally distinct in many important ways (see also Weisberg, 2004, p. 1078). Heterogeneity among target systems, then, causes the trade-offs because greater realism and precision in the model will exclude those systems that either have different causal structures or where those structures vary in degree.

Heterogeneity is especially problematic in population biology, Matthewson argues, because population biologists study populations under natural selection and the first two conditions of natural selection demand difference-making variation. Condition 1 states that there must be phenotypic variation within a population, and condition 2 states that this variation must have consequences for survival and reproduction. The third condition—heritability—ensures that this variation endures. Consequently, population biologists will always face the trade-off that Levins described: either they pursue realism and precision at the cost of generality, or they sacrifice the specificity.

For Matthewson, further evidence that heterogeneity is the main driver of the trade-offs can be found in the absence of any discussion of trade-offs in those disciplines that do not deal with heterogeneous systems (2011, p. 330):

> The fact that trade-offs hold more in population biology than in other natural sciences is evidenced by the fact that although Levins' work influenced many population biologists, his ideas did not noticeably filter through to the other natural sciences… if trade-offs are important and ubiquitous in modelling, then even if physicists or chemists had never heard of Levins or his work, we would expect them to have their own version of "The Strategy" in the relevant literature.

While I think it is undeniable that causal heterogeneity among target systems contributes to the trade-off just described, I wish to make a case for additional trade-offs generated by the complexity of the model structure. In the next section, I will argue that causal heterogeneity does not explain trade-offs in climate science because, at least in the context of climate forecasting, the relevant kind of causal heterogeneity is absent.

2.3 An argument against heterogeneity

My response to Matthewson will begin with an argument that causal heterogeneity is not a good explanation of trade-offs in climate science because the relevant kind of heterogeneity is often missing. In many cases, climate scientists are interested in one target system: The Earth's climate. Moreover, climate modellers often aim to describe that system for a brief window of time: between the years 1850 and 2100.[17] One possible response here is that, although high-fidelity climate models

[17] This claim is based on what can be found in the Intergovernmental Panel on Climate Change's Summary for Policymakers of *The Physical Science Basis* (Stocker, 2014), a report which focuses specifically on 20th and 21st Century climate change. Of course, many climate modellers are interested in different time periods. Paleoclimatologists, most obviously, investigate phenomena in the distant past, far outside of the roughly 350-

typically investigate one system—the terrestrial climate—some need for generality remains. Let's take a look at that response.

Alkistis Elliott-Graves (2018, p. 1109) distinguishes between two kinds of heterogeneity. Intersystem heterogeneity involves variation across a number of different systems, and intrasystem heterogeneity involves variation within a single system across time. Although there is only one actual terrestrial climate, there are multiple possible climate scenarios, which may be enough to create problematic intersystem heterogeneity. For example, modellers want to investigate the Earth where everyone slowly stops emitting greenhouse gases by 2050 and the Earth where everyone continues to emit as usual beyond this point. Representing different emissions scenarios does not create the kind of intersystem heterogeneity Matthewson describes, however, because a single brute force model—that is, a single simulation program—can represent these different scenarios by manipulating the one model rather than by building different models.

A convincing case for problematic intrasystem heterogeneity in climate science is likewise lacking. Although the causal structure of the climate system could change in principle—humans could, for instance, introduce a new artificial component into the climate system—the changing climate is, for the most part, a matter of the climate system and its components occupying different states. It is not a matter of structural change as you might see in ecology when an invasive species is introduced into an ecosystem. While some of these state changes have big consequences for the rest of the state of the system, as in the case of climate tipping points, this should still be conceived of as a change to the system's state rather than the system's structure. Just as with different emissions scenarios, a single high-fidelity model, if it correctly represents the underlying physical processes, can represent both a pre- and post-warming Earth.

To the extent that climate modellers grapple with intrasystem causal heterogeneity, the problem is caused by empirical and computational limitations, rather than, as it in ecology, the very

---

year window that is most relevant to anthropogenic climate change. For the purposes of this paper, however, I am focusing only on one (very prominent!) branch of climate science.

nature of the subject matter. By this, I mean that current knowledge and computational limits may cause modellers to construct their representations such that they appear to be good representations of the climate system given the data we have available but actually introduce biases that limit the generalisability of the model into the future years to which the model is also supposed to apply. Here are two examples of techniques that are used which might cause such problem. Parameterisation schemes are often used to describe poorly understood processes top-down because we do not have the technology to resolve those processes bottom-up from better understood underlying physics. An example of a common parameterisation scheme is one that represents cloud formation, which occurs at relatively fine spatial scales. If cloud behaviour makes a big different to overall climate behaviour, and there is reason to think that it does, then parameterisation schemes biased to current conditions as a result of being built with historical data may not generalise well to the future. Second, ESMs and GCMs are often tuned or calibrated against available data sets. This process of adjustment is intended to achieve a better fit between model and data and make the model a more reliable guide to future climate behaviour. However, again, it's possible that this tuning and calibration can make the model biased toward current and past conditions and, for example, underestimate the impacts of increased warming. The concern that models are biased toward historical conditions is a real possibility that would limit the generalisability of models within the range of their target systems. However, this is an epistemic issue rather than a deep principled problem as it is ecology, where even a perfectly unbiased model of one target will not generalise to another because the two targets are sufficiently structurally dissimilar.

So far, I have argued that there is insufficient causal heterogeneity in the climate science case to generate trade-offs in the way that Matthewson describes. If there are no trade-offs in climate science, then this vindicates Matthewson's argument. If there are trade-offs in climate science, however, then this is a strike against his argument. Based on close textual analysis of Levins' work (1966, 1968b, 1973), Odenbaugh (2003, 2006) argues that Levins' paper was primarily motivated not by target heterogeneity, but by the epistemic problems facing the brute force approach of systems ecology. On this view, it may be possible to build a highly realistic, general, and precise model in

principle, but it would be impossible in practice given the empirical, computational, and cognitive limitations that constrain real modellers. In the next section, I will argue that high-fidelity climate models like ESMs are instances of the brute force approach before arguing, in section 2.5, that they face modelling trade-offs as a consequence of being brute force models. This will support Odenbaugh's interpretation of the trade-offs over Matthewson's.

2.4 Brute force

The "brute force" approach of systems ecology, which Levins criticised, was to build complex computational models including as much causal detail about a target as possible (Watt, 1962; Watt & Watt, 1968). At least at first glance, the brute force approach appears to characterise the high-fidelity modelling found in climate science today.

Recall that the primary component of ESMs and GCMs is a dynamical core, which numerically approximates the "governing equations," a set of seven or eight (seven for the atmosphere and eight for the ocean) non-linear differential equations from the laws of fluid dynamics, to represent the circulation of the atmosphere and oceans (Bjerknes, 1904; Edwards, 2010; Washington & Parkinson, 2005, p. 49). ESMs and GCMs use a 3D grid that represents the atmosphere, the oceans, and land of the entire planet. Using this grid, the computer steps through discretised versions of the governing equations. In general, the finer the resolution the better, but finer resolution comes at the cost of computation time, which places a cap on just how closely the governing equations can be approximated.

In addition to the dynamical core, ESMs contain other components that collectively constitute the model physics, representing other aspects of the climate system, such as sea ice and vegetation. The governing equations, which primarily come from fluid dynamics, have little to say about the rates at which grasslands and forests store and release carbon, or the rate at which they reflect sunlight compared to bare earth. In addition to the modules representing these extra processes, separate

mathematical structures, called parameterisations, represent processes that occur at scales too small to be resolved at current grid resolution.

ESMs and GCMs are examples of what Levins characterised as the brute force approach to modelling. Including as much detail as possible is not a bad strategy if the details matter and in ways that may surprise us due to feedback loops and the connectedness of the system, or if the details are relevant for the kinds of questions we wish to answer with our models (Shukla et al., 2010). However, ESMs and GCMs face a notorious amount of uncertainty largely because of the three problems that Levins attributed to the brute force approach in his classic paper (1966, p. 241):

(a) there are too many parameters to measure; some are still only vaguely defined; many would require a lifetime for their measurement.
(b) The equations are insoluble analytically and exceed the capacity of even good computers.
(c) Even if soluble, the result expressed in the form of quotients of sums of products of parameters would have no meaning for us.

I refer to these kinds of problem as (a) the problem of data hunger, (b) the problem of tractability, and (c) the problem of comprehensibility. In the remainder of this section, I will demonstrate that high-fidelity climate models face these problems and that practices used to manage these problems can be sources of uncertainty unto themselves.

2.4.1 The problem of data hunger

The first problem Levins identified with highly detailed simulation models is that they require a large amount of data to fill-in the details. Measurements are needed to adjust parameters, set initial conditions,[18] and evaluate model performance. Climate science is a very data hungry discipline indeed.

[18] In some conditions, models do not rely on specific initial conditions. ESMs and GCMs, in fact, are often allowed to "spin-up" for some simulation time, falling into their own natural equilibria before forcing scenarios, such as different carbon emission schemes, are imposed and the model system moves away from its equilibrium.

As Edwards argues in his (2010), climate science as a discipline is predicated on the fairly recent global infrastructure of weather stations along with the more advanced technology required to take measurements beyond the Earth's surface, such as weather balloons and satellites. As Levins argued, data hunger becomes particularly problematic when the relevant data is difficult or impossible to collect, as it is in climate science today. This is to say nothing of the technological requirements necessary for storing and sharing the large amounts of weather data needed to describe the climate.

For example, one way to evaluate how well a model represents Earth's climate is to compare the model's output to real-world data (Winsberg, 2018a). To do that, we need some data, and while our international system of weather and climate measurement is better today than it was on Balloon Day 1910, our weather stations, ocean buoys, weather balloons, and so on, are not distributed evenly across the Earth's surface, up into the atmosphere, or down into the oceans and soil. Certainly not with a density proportional to the 3D finite different grid of an ESM. These grids can have points every 100km across the horizontal of the Earth's surface and at far smaller intervals than that along the vertical up into the atmosphere and down into the ocean. We simply do not have this much equipment. Moreover, as the resolution of climate models increases, so does their appetite for data.

Scientists could, in principle, cover the world in measuring equipment going into some science fiction future—we have smart homes, there is a large literature on smart cities (see Meijer & Bolívar, 2016), so why not a smart Earth? Unfortunately, there would be no way to deploy such information infrastructure on any past Earth. The extent to which Climate records of the 19th and 20th century can grow is heavily constrained, but this is a key period used to evaluate model performance: "data are available for only a few quantities (e.g. temperature, pressure, precipitation), for only relatively recent time periods, and primarily for land locations and near-surface locations, and even these records are incomplete and of variable quality" (W. S. Parker, 2006, pp. 353–354). Historical data only gets patchier the further back in time we go.

Scientists, ever ingenious, have developed techniques to fill the gaps in the historical records by calculating the likely values for the missing data points. Data sets produced through such a process

are known as reanalysis data sets (W. S. Parker, 2016). Reanalysis, however, is its own source of uncertainty. Reanalysis data sets are constructed using weather models that are not identical to ESMs, but that have a shared history and are based on numerical approximations of the same physical equations representing fluid flow. This similarity introduces epistemic complications when evaluating a model against a reanalysis data set (Lewis 2017). While a fit between model *M* and data is typically a good sign, fit between *M* and a data set partly constructed using algorithms resembling those numerically approximating the same equations in *M*—as in the case of reanalysis—provides comparatively less confidence about the quality of *M*. The similar data and model output can be explained by appeal to a common causal factor—approximation of the same dynamical equations—rather than by appeal to the model's skill at reproducing natural patterns.

2.4.2 The problem of tractability

Levins argued that highly complex models are limited by computational power. Running a simulation that included every relevant causal detail about some ecosystem was simply not feasible given 1960s technology. Causal completeness still isn't feasible in climate science today (Winsberg, 2018a). The ideal model, then, cannot be investigated until we make more powerful computers. Until then, we are stuck with non-ideal models that are made with rough approximations, ad hoc adjustments, and parameterisations.

For example, Bjerknes' governing equations are based on well-established pieces of physical theory, which we might trust to represent atmospheric and oceanic flows, but these equations must be numerically approximated and finer approximations require more computational power, which we do not have. Approximations are made rougher through truncation and rounding errors resulting from memory and computational limitations. The result of some calculation at one grid point may have a long string of numbers after the decimal point, not all of which can be stored in the computer's memory, so the numbers after a certain position, such as the fourth decimal place, are simply dropped and forgotten (truncation), or dropped and forgotten after rounding the last remaining digit up or down

(rounding). Although a single instance of truncation or rounding may make little difference, these little differences can accumulate as the simulations run over long time scales and the result of one calculation is used as the input for the next (Rosinski & Williamson, 1997).

Computational limitations also force modellers to use parameterisations to represent important processes that occur at scales too small to be resolved within an ESM's finite difference grid. For example, shallow cloud formation in the lower atmosphere is a major contributor to Earth's albedo factor—that is, how much incoming solar radiation is reflected out of the system and doesn't contribute to the Earth's energy budget—and consequently to the Earth's climate. Interactions with aerosols can also determine droplet size and alter the cloud's albedo (Pasini, 2005, pp. 126–127). Cloud behaviour, however, occurs on scales that even the finer grids on the market cannot resolve (T. Schneider et al., 2017, p. 4):

> atmosphere models have a horizontal grid spacing around 50-100 km and a vertical grid spacing in the lower atmosphere around 200 m. This is much too coarse to resolve the 10-100 m wide turbulent updrafts that originate in the planetary boundary layer and generate low clouds.

Tapio Schneider and colleagues calculate that, given current grid spacing, the improvements in grid spacing required to resolve low clouds, and the pace at which computational power advances, sufficiently fine grids will not be available until the 2060s. Until then, shallow cloud formation must be represented separately in the form of a parameterisation scheme. However, we do not yet understand this process or how best to model it (Stocker 2014), and different cloud parameterisation schemes account for the bulk of the disagreement among climate model projections (T. Schneider et al., 2017, p. 4). As stated in the Intergovernmental Panel on Climate Change's Fifth Assessment Report: "There is *very high confidence* that uncertainties in cloud processes explain much of the spread in modelled climate sensitivity" (Stocker, 2014, p. 743).

In response to parameterisation use and computational limitations, a model's equations may be adjusted in ad hoc ways (Edwards, 2010; Lenhard & Winsberg, 2010; W. S. Parker, 2006;

Winsberg, 2010). This can occur during a process known as model tuning: “Tuning a climate model involves making ad hoc changes to its parameter values or to the form of its equations in order to improve the fit between the model’s output and observational/reanalysis data” (W. S. Parker, 2011, p. 587). Rough approximation, parameterisation, and ad hoc adjustment are responses to computational limitations, which obscure the connection between the model’s behaviour and the solution to the theoretically principled equations. This is a problem if we rely on the credentials of the governing equations to justify the inferences we make with our model.

2.4.3 The problem of comprehensibility

Levins argued that highly complex computational models are incomprehensible to human scientists. Complex climate models are millions of lines of code long, typically scripted by teams of scientists, and run on powerful supercomputers. This contrasts with models described with only a couple of equations that can be investigated using back-of-the-envelope reasoning and which might characterise Levins’ simple theory approach. Compared with simple models, the inner workings of ESMs and GCMs are too complicated, and the output a number array too large, for any human scientist to wrap their head around.

To some extent, the problem of understanding complex simulations has been addressed with computer-assisted analysis techniques which take number arrays and convert them into some form that is comprehensible to a human. Most notably in computational modelling more generally, visualisation techniques can be used to present the values of variables over time in a fashion that appeals to intuitive perception. Some philosophers have even argued in the past that visualisation is an essential aspect of computer simulation (Hughes, 1999; Humphreys, 2004, pp. 112–114; Rohrlich, 1990, p. 515). As Eric Winsberg argues, visualisation appealing to intuitive perception is just one technique among many for analysing simulation behaviour, but such computer-assisted techniques are required to make sense of the large number arrays produced by a complex computational model in just the same way that they are required for making sense of large data sets (Winsberg, 2010, p. 33).

While techniques like visualisation assist with presenting the results and behaviour of climate models in a way that human modellers can comprehend, a deeper problem remains. Representing the key components of the climate system together in one comprehensive model can obscure causal attribution within the model: “Interpretation of cause and effect linkages may be difficult to trace in a GCM because of the large number of internal degrees of freedom in the model and because of the huge volume of output generated by a high-resolution time-dependent model” (S. H. Schneider & Dickinson, 1974, p. 486). This is a problem because, if modellers struggle to understand why their models produce the results and behaviours they do, then it becomes more difficult to improve these models or assess their limitations.

For example, a complex model structure comprised of multiple sub-modules and parameterisation schemes representing different climate processes can also create the possibility of compensating errors (Winsberg, 2010, 2018a, pp. 196–197). That is, when a model gets something wrong (or right, for that matter), we won’t know where to place the blame. A result that seemingly vindicates the model—say, a match between the model’s performance and historical data—could be caused by errors in model components that compensate for one another when coupled. Although introducing two errors into a model that miraculously cancel each out might initially seem unlikely, it is more common because of the practice of model tuning. Model tuning is a process where the parameters of the different modules and parameterisation schemes are adjusted to increase the similarity between the model output and observed data. In this situation, compensating errors are not unlikely to emerge because the adjustments made to model components are evaluated according to the model’s performance as a whole. Compensating errors can become a real problem when modellers move to future climate scenarios that are radically different from past or present climates. Here, the errors that had once cancelled each other out may no longer do so, leaving us with a model that appears accurate given the data we do have, but is nevertheless wildly inaccurate for the future cases that matter for large-scale decision-making and planning.

2.4.4 Summary

Let's recap. Levins identified three problems for the brute force approach, which I have applied to contemporary climate modelling. First, data hunger is a problem for highly complex models like ESMs. The success and existence of climate science requires a global information infrastructure, but this infrastructure is limited by time and money going forward into the future and by the lack of technology and infrastructure looking back into the past. The best attempts to patch gaps in historical data using reanalysis is a further source of uncertainty as the algorithms used in climate models like ESMs are based on the same mathematics and even have a shared history with those used in reanalysis. Second, the equations representing the underlying physics of the model must be approximated due to computational limits, and grid sizes are constrained to a scale that does not permit the resolution of important sub-grid processes that play crucial roles in the climate's overall behaviour. And third, errors within a model can be obscured by the model's complexity and multiple interacting components can produce a confluence of errors which cause poor performance when projecting to future climate states very unlike the historical conditions to which the model has been tuned, but which may go unnoticed when the models are compared with historical data.

Next, I will draw the threads of the last few sections together to argue that these three epistemic challenges facing brute force models generate the modelling trade-offs in climate science rather than causal heterogeneity among target systems.

2.5 The trade-offs in climate science

Although Matthewson might be right that target heterogeneity is the primary source of modelling trade-offs in population biology, Odenbaugh's view that the trade-offs are generated by pragmatic empirical, computational, and cognitive constraints, is a natural fit for climate science. This is both because, as I argued in section 2.3, many high-fidelity climate models represent the one target system, and, as I argued in section 2.4, these models butt up against the empirical, computational, and cognitive constraints that Levins attributed to brute force models.

Indeed, Odenbaugh's line of argument may apply better to climate science than it does in population biology. According to Matthewson, Odenbaugh's response undersells the importance of Levins' analysis because it focuses on contingent factors rather than inescapable features of the logic of representation (Matthewson & Weisberg, 2009) or the subject matter of the discipline (Matthewson, 2011, p. 328):

> We have presumably already overcome many of the practical limitations that existed for Levins and his peers in 1966… the more that Odenbaugh convinces us that these trade-offs are only due to contingent limitations, the less compelling Levins' claims become.

Providing grist to Matthewson's mill, Odenbaugh (2006) agrees that the epistemic challenges Levins describes are not as striking in contemporary biology as they once were. Rather, new computational and mathematical techniques, such as agent-based modelling and matrix algebra, have at least partially addressed the problems of tractability and comprehensibility. The measurement problem remains however, and, on his view, has always been "the most serious problem for population biology" (Odenbaugh, 2006, p. 620). However, even if Odenbaugh is right and these pragmatic problems have been partially addressed in ecology, they persist in climate science and it is unlikely that modellers will overcome these limitations any time soon (T. Schneider et al., 2017, p. 4). Indeed, the incomprehensibility and causal opacity of ESMs may only get worse with technological improvements if climate modellers include yet more components in their state-of-the-art models as the technology permits.

In the remainder of this section, I will make one final argument against Matthewson, focusing on his (2011, p. 330) claim that, if a scientific discipline faced a trade-off, it would have its own literature on an equivalent of "The Strategy." As we will see, climate modelling does, in fact, have such a literature despite not facing a problem of heterogeneity like population biology, further demonstrating that model and target complexity alone is sufficient to generate trade-offs.

Jule Charney, an atmospheric modeller from the early post World War II days and beyond, argued that climate modellers, or anyone representing similar complex systems, must "choose either a

precise model in order to predict or an extreme simplification in order to understand" (Charney, 1963, p. 289; c.f. Dalmedico, 2001, p. 415).[19] The central trade-off in climate science, then, is between realistic and precise models and comprehensible ones.[20] Highly realistic models have their strengths: representing as much causal structure as possible will hopefully help avoid the problem of omitting a potential difference-maker or missing some unforeseen feedback loop, and more detailed models can be used to investigate more detailed counterfactuals, such as the change in Australasian precipitation patterns in response to three degrees of warming (Shukla et al., 2010). As Charney suggested, these strengths make them well-suited to the task of prediction.

Simpler models, on the other hand, are far more comprehensible than complex ones and, as such, are far better suited to fostering understanding of fundamental climate processes. Most obviously, simpler models are better suited to isolating key difference-making processes in contrast to realistic or comprehensive models that try to include as many processes as possible. Simpler models are also better suited to increasing understanding within or across an epistemic community because simpler models can have greater longevity. While realistic models may become obsolete as more powerful computers become available and the state-of-the-art changes, simpler models retain their value precisely because their value is not, for the most part, determined by the limits of technological possibility. The longevity of simple models, Nadir Jeevanjee and colleagues (2017) argue, could foster a greater understanding of fundamental climate processes because a smaller set of simpler models lends itself to thorough investigation by researchers over time. A large set of models hidden away in

[19] Note that Charney is not using "precise" in the way that it has been defined in section 2.1. A model with finely specified parameters and variables *can* still be incredibly simple. An EBM is such a model. Rather, given that Charney is discussing models that were the state-of-the-art in his day and from which today's ESMs descend, we can read his comment as one about the trade-off between realistic and precise models directed toward prediction and simple models directed toward understanding.

[20] Levins (1993, p. 554) acknowledged that understandability was another important modelling desiderata beyond the three that his (1966) arguments focused on.

many different publications, they argue, is less apt to foster such collective comprehension and progress.

In addition to an old and on-going literature on its own version of the strategy, climate scientists have discussed ways of conceptualising the space of different models within this literature. Charney, for example, described a process of climbing a "hierarchy of models" in which climate models would slowly become increasingly comprehensive until researchers reached the most realistic model at the top of the hierarchy (Edwards, 2010). Modellers were to achieve greater understanding of climate processes as they moved up and built this hierarchy. Unfortunately, these hierarchies were somewhat forgotten for several decades while modellers pursued the alternative brute force approach of building ensembles of highly complex models—a trend facilitated by the rapid development of computational hardware (W. S. Parker, 2014). As modellers have continued to wrestle with the epistemic challenges of this approach, however, there has been some renewed interest in model hierarchies:[21]

> Solid progress toward an understanding of the dominant factors in climate change will require steady development of an almost continuous spectrum or hierarchy of models of increasing physical and mathematical complexity. (S. H. Schneider & Dickinson, 1974, p. 486)

> The models used to simulate the climate are themselves complex, chaotic dynamical systems. To work with them effectively requires not only the careful examination of alternative formulations of these comprehensive models but also the construction of a hierarchy of models in which elements of complexity are added sequentially. (Held, 2014, p. 1206)

> One major benefit of model hierarchies is in providing simplified versions of systems of interest, which are easier to study and generate hypotheses about… Conversely, hierarchies

[21] As I will note below, talk of hierarchies can be misleading, and these are much better thought of as spaces than hierarchies.

> can also tell us that things that are difficult to understand in the comprehensive system may remain so even in the simplified systems. (Jeevanjee et al., 2017, p. 1763)

There have been a few different ways of conceptualising of the dimensions of the model hierarchy (e.g. Held, 2005; Jeevanjee et al., 2017, p. 1762; McGuffie & Henderson-Sellers, 2013, p. 52). In recent papers, Sandrine Bony et al. (2013), Nadir Jeevanjee et al. (Jeevanjee et al., 2017), and Penelope Maher et al. (2019) have all discussed model hierarchies in climate science, with Maher et al. focusing specifically on atmospheric models. In this chapter, I will only present Bony et al's conceptualisation since it is the simplest. But we will be returning to Jeevanjee et al.'s conceptualisation of the model hierarchy in Chapter 3.

My reason for ending this chapter with a presentation of an example model hierarchy is to emphasise both that climate scientists face modelling trade-offs in virtue of model and target system complexity and that they have a literature on the topic. This runs counter to Matthewson's claim that disciplines avoiding the problem of heterogeneity do not face the trade-offs or, at least, face them less severely, and do not have a literature on the subject. At this point, I should also note that, as Jeevanjee et al. claim, the notion of a hierarchy can be misleading, since there is not a strict ordering of less to more complex models: "how, for instance, can one compare a moist, non-rotating cloud-resolving simulation in a planar geometry to a dry, rotating, coarse-resolution global simulation? One is not clearly more realistic than the other, at least in any general sense" (Jeevanjee et al., 2017, p. 1761). Henceforth, then, I will break with the naming convention of climate science and refer to these as "model spaces."

Bony et al. represent climate model space with two dimensions, as shown in figure 2.1. What Bony et al. call "system complexity" runs along the x axis, spanning basic particle and fluid systems at one end of the spectrum, to the entire Earth and Earth-Human systems at the other. Recall, from Chapter 1, the distinction between phenomena and target systems, with target systems being parts and properties isolated from phenomena. All the target systems positioned along the x axis are taken from the same phenomena—that is, the Earth, its climate, and the things living on the Earth that contribute

to the climate—but isolate different parts and properties from this phenomena. Isolating the Earth and human systems results in a more complex target system than if one omits the details of the human parts and properties. As Bony et al. note, less complex systems are typically more amenable to experimental isolation and investigation in laboratory conditions, while this is frequently impossible for more complex systems like the Earth system or the Earth-Economy system.

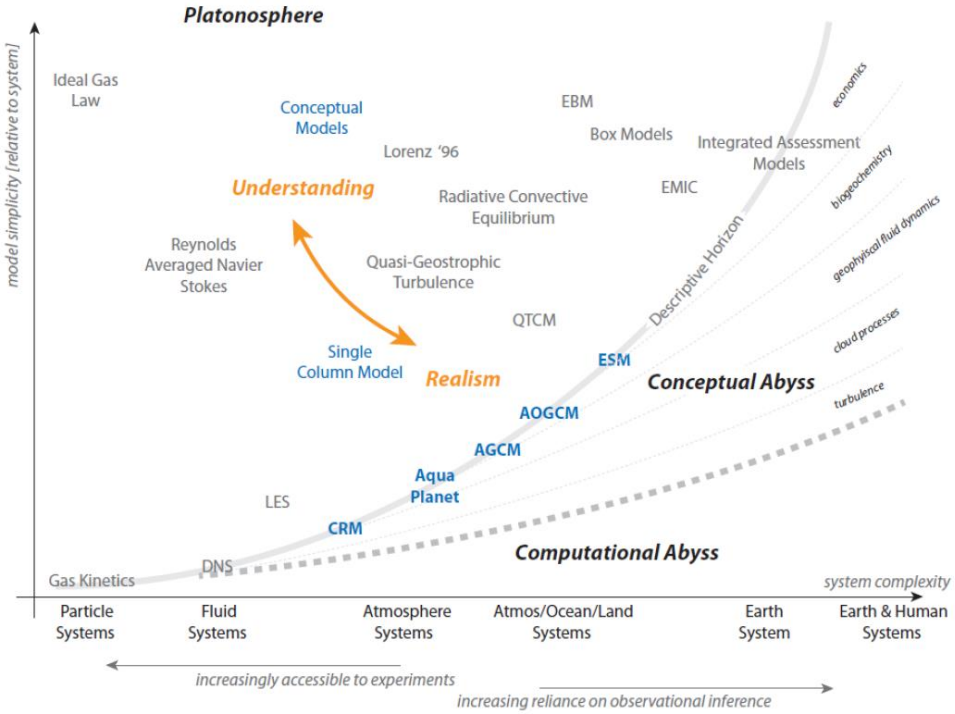

Figure 2.1 Bony et al.'s two-dimensional space of climate models. The space shows the trade-off between understanding and realism associated with choosing to build simple models of simple targets or complex models of complex targets. It also demonstrates that realistic models are constrained both by what is feasible given the computational resources required to build complex models and by what is possible given what we know (and do not know) about the target (reproduced with permission from Bony et al., 2013).

Model simplicity relative to target system runs along the y axis. Bony et al.'s figure shows that EBMs and ESMs are directed toward a target system of near-identical complexity: The Earth system. However, EBMs are far simpler relative to that target than ESMs, with the latter resulting from an attempt to include as much detail about the target given limitations in computational power and

empirical knowledge.[22] In figure 2.1, Bony et al. show the trade-off between comprehension and comprehensiveness with a yellow arrow running from the top-left corner (simple targets, simple models, good for understanding) down to the bottom-right corner (complex models, complex targets, poor for understanding).

Bony et al. label three regions in their space. First, they label the top-left corner the Platonosphere because it is populated by models that are both very simple and have targets that are highly abstracted patterns, hence the reference to Plato's forms. In virtue of their simplicity, it is the representations populating this and nearby regions that we understand the best. In addition to the Platonosphere, Bony et al. label two other regions of their space the "conceptual abyss" and the "computational abyss." The computational abyss represents model space that we cannot access due to the limitations of our computing power. Perhaps we could build these models, but we would not be able to investigate them in any reasonable amount of time. The exact boundaries of the computational abyss will shift as advances in computing technology are made. The conceptual abyss, on the other hand, is a region that is unpopulated not because these models would be computational intractable, but because we do not understand the processes operating within the targets sufficiently to include them in our representations. While modelling poorly understood processes is common and an important way of exploring hypotheses, this is typically performed with much simpler models because they are more

[22] Bony et al.'s approach to categorising models raises a question about target systems and models. Can a model be far simpler than it's target system, omitting many of the parts and properties that the modeller has deemed sufficiently important to include in the target system, or must a model contain all parts and properties included in the target system? In my view, models need not contain all the parts and properties of the target system. This allows, for example, cases in which modellers wish to represent their target with multiple simpler models (See Weisberg, 2013, Chapter 6 for a description of such cases). Similarly, the modeller may believe that a set of parts and properties are relevant to a focal pattern they wish to recreate or explore, so rightly belong in the target, but must be omitted from the model for the sake of simplicity or tractability. It is then the job of the model construal, and particularly the fidelity criteria, to state which parts and properties of the target it is acceptable for the modeller to omit from the model relative to their purpose.

transparent, isolating the particular hypotheses being modelled. There is little value introducing these poorly understood processes into highly complex models aiming toward prediction as the guess work required to include the processes can undo the good work done by the representations of well-understood processes. The limitations of the computational and conceptual abyss create what Bony et al. call the "descriptive horizon," which marks the limit of useful and tractable models.

Before we leave Bony et al.'s particular conceptualisation of the model space aside, let me say one last thing about the "conceptual abyss." Pointing out that our highly complex models are constrained by what we know about the targets is a fourth epistemic challenge facing brute force models, which was not recognised explicitly by Levins. While Levins recognised that brute force models are difficult to understand, understanding the target system is just as important. Likewise, modellers would still need to know what kinds of data they ought to be collecting and using to improve their model even if data collection was no obstacle. This further demonstrates the need for simpler models, which can assist in the generation of knowledge about the target. This knowledge can then feed back into the process of model construction when it comes to brute force models.

Bony et al.'s model space is apt for demonstrating the trade-off between comprehension and comprehensiveness as well as the pragmatic limitations that high-fidelity modellers face. For now, I will restate that my purpose in presenting Bony et al.'s model space has been to demonstrate that climate modellers have an on-going literature on modellingh trade-offs and that these trade-offs are generated by model and target complexity rather than target heterogeneity. From this discussion, it should be clear that brute force modelling comes with a price and should not be pursued in isolation and in the absence of simpler models directed toward understanding general patterns rather than predicting the behaviour of specific systems.

2.6 Conclusion

In this chapter, I hope to have shown that Levins' "The Strategy" is a seminal entry in the philosophy of computational modelling that remains relevant today. Written at a time when simulation science

was just breaking into population biology, Levins' articulation of the epistemic challenges faced by the brute force approach remains pertinent to climate science where the strategy flourishes. As Levins described, models with highly complex structures intended to represent their targets with a high degree of realism and detail have big appetites for sometimes unattainable data, are intractable in their ideal forms, so must be adjusted in various ways to run on available computers, and their complexity makes them difficult to understand. These epistemic problems create trade-offs and imply that, although computer simulation has enabled modellers to investigate model systems that they otherwise could not, there remains an important place for simpler models intended to foster understanding of key processes.[23]

With respect to the philosophy of science literature on modelling trade-offs and the debate between Matthewson and Odenbaugh, I have shown that, even if target heterogeneity generates modelling trade-offs, complex model structures and the epistemic challenges they face are sufficient to generate an important set of trade-offs to which model pluralism is the best response. Simpler models do not require as much, potentially unobtainable, data as complex models. Simpler models do not require the same intense computational resources as more complex models. And simpler models are more transparent and better at fostering understanding than more complex ones. The arguments from this chapter leave us with the following lesson for the epistemology of modelling and computer simulation: Comprehensive representation of a complex target comes at the cost of comprehensibility.

The discussion of this chapter leaves at least two questions unanswered. First, what exactly is this scientific understanding that I have mentioned and how exactly do simpler models achieve the goal of fostering understanding? Second, for those modellers pursuing the brute force approach in earnest, what can be done to mitigate the three epistemic challenges? Answering these questions will take us to a discussion of the second highly influential part of Levins' paper, which was his description of a method called "robustness analysis." As I will argue in Chapter 3, robustness analysis and

[23] It should be noted that these simpler models are often still sufficiently complex or non-linear that their investigation requires computer simulation.

fostering scientific understanding go hand in hand. The second question will remain unanswered until Chapter 6.

3

The strategy of model building in climate science II: Robustness analysis, model hierarchies, and scientific understanding

In this chapter, I continue my application of Levins' work to contemporary climate science. Where the previous chapter focused on modelling trade-offs, this chapter focuses on robustness analysis, which Levins described as a method for evaluating the impact of idealising assumptions of model results. In this chapter, I argue that philosophers of science have overlooked an application of the concept of robustness analysis to the practices of climate science. While most of the attention has been on ensembles of highly complex climate models, I demonstrate that climate sciences model hierarchies, better conceptualised as model spaces, can be seen as a means of using simpler models to identify key causal relationships. I also argue that robustness analysis and scientific understanding are closely connected, with robustness analysis being a method for examining causal dependencies and scientific understanding being achieved when a scientist grasps a causal dependency.

3.1 Introduction[24]

The previous chapter applied Levins' (1966) 'The Strategy' to contemporary climate modelling, arguing that comprehensive representation comes at the cost of comprehensibility. This chapter continues to mine Levins' work in the context of climate modelling, exploring Levins' notion of *robustness analysis* and attempting to find some clarity regarding the aim of *scientific understanding* that I introduced in the previous chapter without defining. Ultimately, I will argue that robustness

[24] Some of the material included in this chapter has been published in (Walmsley, 2020).

analysis and scientific understanding are closely related and that this connection has largely been missed by much of the discussion of robustness analysis in the philosophy of climate science.

This chapter proceeds as follows. In section 3.2, I will begin by describing Levins' robustness analysis. At least as it was initially proposed, robustness analysis appeared to be a method for increasing a modeller's confidence that the result of their model was due to it successfully depicting a real causal relationship rather than a result of the specific idealisations used in the model's construction. Next, in section 3.3, I will present part of the discussion of robustness in the philosophy of climate science. Within the philosophy of climate science, robustness analysis has been discussed at length (e.g. Lehtinen, 2016, 2018; Lloyd, 2010, 2015; Parker, 2011, 2013; Winsberg, 2018a, 2018b). A prominent view is that ensemble modelling (a procedure described briefly in section 1.2.2 and in greater detail below) does little to reduce model uncertainty, so does not count as an instance of robustness analysis. This is because these ensembles represent a very small sample of possible model space, the models are not varied systematically, and the models are not independent (for a more optimistic view see Lehtinen, 2016, 2018; see also Weisberg, 2006b).

In section 3.4, I will argue that ensemble modelling is not the only practice in climate science to which robustness analysis may be applied. Instead, by applying Tarja Knuuttila and Andrea Loettgers' (2011) notion of causal isolation robustness to model spaces (what climate modellers called a "model hierarchy" in the previous chapter), I will claim that we can consider robustness analysis as a practice of comparing complex models with simpler ones, progressing systematically through model spaces to assist modellers to identify and articulate possible causal dependencies. In section 3.5, I will present a view of scientific understanding which is compatible with my conception of robustness analysis, demonstrating how understanding is a valuable scientific goal that can be achieved with the assistance of systematic exploration of model spaces. My aim in making this argument is not only to contribute to these debates within the philosophy of climate science but to demonstrate how the notion of a model space is beneficial to the general literature on robustness analysis, serving as a template for systematic robustness analysis. Finally, in section 3.6, I will conclude by summarising the lessons of this chapter for the epistemology of models and computer simulations more broadly.

3.2 Levins' robustness

Robustness means many things to many people (Lisciandra, 2017); another promiscuous term like "model." In addition to Levins' original framework, William Wimsatt (1981, 2007) uses a range of robustness concepts, which Brett Calcott (2011) narrows down to three. James Woodward (2006) has his own way of drawing the distinctions, and Michael Weisberg and Kenneth Reisman (2008) extend Levins' framework with three kinds of robustness analysis. Elisabeth Lloyd (2010, 2015) has advanced her own notion of model robustness especially with climate models in mind, and Jonah Schupbach (2016) has proposed explanatory robustness as a unifying robustness concept, which Eric Winsberg (2018a) endorses as the best account of robustness in the context of climate science. We will touch on some of these distinctions in detail later in this chapter where they are useful for our discussion. For now, I will leave a brief description of these different robustness concepts in table 3.1 and provide an intuitive gloss here which captures most of what these concepts have in common. At its core, robustness relates to the stability of a result, often in the context of consulting multiple sources to draw an inference. This might range from using multiple models to find a result, using multiple instruments to take a measurement, or even questioning multiple witnesses to get an account of an event, but can also refer to the invariance of a phenomena to a range of different conditions. While I will return to the other robustness concepts, my aim in this section is to describe Levins' original notion of robustness analysis before turning, in section 3.3, to the details how Levins' idea has been discussed and debated within the philosophy of climate science, especially in the context of ensemble modelling.

Author(s)	**Robustness concept**	**Brief description**
	Robust theorems	A model result can be derived from multiple models.

Wimsatt (via Calcott p. 283-4)	Robust detection	A claim about the world can be detected through multiple independent means, such sensory modalities or experimental procedures.
	Robust phenomena	A phenomenon is invariant under a range of contexts.
Woodward (2006)	Inferential robustness	A conclusion can be drawn with an inference from a data set and additional assumptions, and there are multiple, competing hypotheses about how those assumptions should be filled in.
	Derivational robustness	A model result can be produced even if assumptions (such as the value of a parameter or the relationship between parameters) are varied.
	Measurement robustness	Multiple measurement procedures produce a very similar value for some quantity.
	Causal robustness	A causal relationship is invariant to range of interventions or manipulations.
Weisberg and Reisman (2008)	Parametric robustness	A model result is robust with respect to a parameter if the result can be derived despite changes to the values of that parameters.
	Structural robustness	A model result is robust with respect to its structure if the result can be derived despite changes to the processes included in the model structure.
	Representational robustness	A model result is robust with respect to its representational mode if the result can be derived despite changes to the way in which the model is represented

		(such as a change from an equation-based to an agent-based model).
Lloyd (2010, 2015)	Model robustness	A model result is robust if it is derived from multiple independent models where the assumptions of each are empirically supported.
Schupbach (2018)	Explanatory robustness	A hypothesis is robust to the extent that alternative, competing hypotheses have been ruled out.
Knuuttila and Loettgers (2011)	Independent determination robustness	Independent determination robustness involves increasing researchers' confidence regarding a result by using multiple lines of evidence
	Causal isolation robustness	Causal isolation robustness involves determining the sufficient conditions required to produce a target phenomenon.

Table 3.1 A non-exhaustive list of robustness concepts from the philosophical literature.

Let's start with Levins' original notion. According to Levins, robustness analysis involves building a family F of models M_i with a fixed causal core C and a set of varying auxiliary assumptions A_i. If, despite the variety of A_i, the models produce the same property R, then the relationship between C and R is robust and modellers can formulate a "robust theorem" describing their relationship (Levins, 1993, p. 553); something like *a general biocide* (the causal core) *favours the relative abundance of the prey* (the result) (Weisberg & Reisman, 2008). Another example might be *increased albedo factor* (the causal core) *reduces global average surface temperatures* (the result). This is shown schematically in figure 3.1.

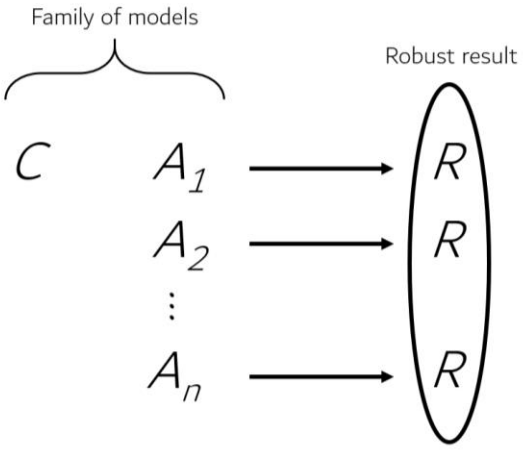

Figure 3.1 Modellers identify a robust property *R* by building a family of models with a fixed causal core *C* and a set of auxiliary assumptions A_i. If R appears in all or a sufficiently large proportion of models in the family, then it is robust and a we articulate a robust theorem stating the relationship between *C* and *R*.

Although Levins did not analyse the structure of robust theorems closely, Weisberg (2006b) does, and offers a helpful formulation of robust theorems as ceteris paribus statements (Weisberg, 2006b, p. 738):

> *Ceteris paribus*, if [common causal structure] obtains, then [robust property] will obtain.

He also states that "a fully formed robust theorem has three parts: a common structure, a robust property, and a set of *ceteris paribus* conditions" (p. 738). Finding the conditions under which an identified *C* to *R* relationship breaks down is a key step in articulating a generalisation. That is, uncovering robust theorems involves a kind of "scope and limit" analysis of the causal dependency described.

Recall that Levins' (1966) paper was a response to the criticisms of the systems ecologists like Watt, who argued that the modelling approach taken by Levins and colleagues produced models that were too idealised to be of any use. In that dialectic context, the procedure of building families of systematically varied models can be seen as a method for ensuring that the simplicity and idealisations

of these models did not interfere with their ability to provide insight into dependencies operating in the real, complex world. Controversially, Levins appeared to suggest that robustness analysis was a kind of confirmation procedure or "method of validation" (Levins, 1968b, p. 7), famously stating that "our truth is the intersection of independent lies" (Levins, 1966, p. 423).[25] According to his own (1993) clarification, however, robustness analysis should not be considered a search for the truth *per se*, and so should not be considered a confirmatory exercise. Rather, it intended to be a strategy for dealing with uncertainty: "The search for robust theorems reflects the strategy of determining how much we can get away with not knowing, and still understand the system" (p. 554). In section 3.5, I will be taking a closer look at this suggestion that robustness analysis serves to produce scientific understanding, comparing it to the view that robustness analysis plays a confirmatory role.

Within the philosophy of science more broadly and the philosophy of climate science more specifically, there has been a large amount of discussion of robustness analysis, its purpose, and what it is able to achieve. To continue the approach of the previous chapter, I will focus on robustness analysis in the context of climate science, reviewing the literature on the subject in section 3.3 before presenting my own view in sections 3.4 and 3.5. Although I will be using climate models as my exemplars, my view of robustness analysis and the methods by which modellers can extract the most value from the practice generalises to all disciplines.

3.3 Robustness analysis in the philosophy of climate science

At least superficially, robustness analysis—that is, improving the epistemic situation by using multiple varied models—appears to be commonplace in climate science. One way in which modellers deal with the uncertainty of ESMs is to perform ensemble studies (Lloyd, 2010; Odenbaugh, 2018a; W. S. Parker, 2006, 2013b). There are two kinds of model ensembles. Perturbed physics ensembles are

[25] Weisberg also takes robustness analysis to provide a kind of low-level confirmation, but there is great disagreement about its confirmatory power (see Forber, 2010; Lehtinen, 2016; Levins, 1993; Lloyd, 2010).

specifically meant to address parametric uncertainty and involve running the same model structure with a range of different parameter settings. Multi-model ensembles, on the other hand, involve using multiple different models from different modelling centres and comparing their behaviour in the same scenarios. Multi-model ensembles are intended to address structural uncertainty by, for example, using different parameterisation schemes for poorly understood but important processes like cloud formation (Stocker 2014).

At first glance, both perturbed physics and multi-model ensembles seem a natural fit for the notion of robustness analysis discussed in the previous section. Note that perturbed physics ensembles really are an example of the parametric robustness analysis described by Weisberg and Reisman (2008) and included in table 3.1. My focus, however, will be on multi-model ensemble studies. Despite the similarities between these ensembles and Levins' robustness, many philosophers have argued that ensemble modelling does not meet the conditions of robustness (Lloyd, 2010, 2015; W. S. Parker, 2006, 2010, 2011, 2013b; Winsberg, 2018a, 2018b), although some are more optimistic (Lehtinen, 2016, 2018). To clarify, the claim is not that one is unable to conceive of ensemble studies as robustness analysis. Rather, the claim is that, if one were to conceive of ensemble studies as an instance of robustness analysis, then they would not appear to be particularly successful instances of it. To explain why ensemble modelling is useful, then, one would need to look beyond robustness analysis.[26]

One reason for thinking that ensemble studies do not meet the conditions for successful robustness analysis are that these ensembles cover a very small region of possible model space, especially considering that the number of adjustable assumptions in highly complex models like ESMs. Naturally, all instances of robustness analysis involve a relatively small investigation of possible model space because they keep the common core fixed and only vary the relevant auxiliary

[26] As Parker has argued, it's unclear exactly what to make of any agreement in a multi-model ensemble. One possibility is that the agreement just reflects the field's "best guess," but not that we should have greater confidence in that guess (W. S. Parker, 2006, 2010, 2013b, 2018).

assumptions used to construct the model. However, in the case of ESMs, the number of auxiliary assumptions that should be varied in a thorough robustness analysis is overwhelming, and adjusting them all would require ensembles far larger than the collection of fifty or so models seen in the ensembles of the last Intergovernmental Report on Climate Change assessment (Stocker, 2014). Even within the small ensembles that are studied, questionable assumptions are not varied one at a time and there may be many structural differences between one research group's model and another's. Unfortunately, this is not the result of carelessness that could be easily fixed. Rather, it is the result of hard constraints. The size of ensembles are effectively constrained by a higher order version of the computational tractability challenge faced by brute force models and described in the previous chapter: it takes a lot of time to investigate one model, let alone a large set of systematically varied models.[27]

To make matters worse, the models within these multi-model ensembles are not independent (Carrier & Lenhard, 2019). Naturally, they are built upon the same background knowledge of climate theory and are benchmarked against the same data. But, more problematically, they also have a common history, with some model components, such as the algorithms that approximate the primitive equations, being shared between research groups (Edwards, 2010). Modellers also move between research groups and may take their methods with them. This further reduces the model space that multi-model ensembles cover. As before, models ought to share a common causal core performing robustness analysis, so they need not be completely independent, but the lack of independence in the present case goes beyond the causal core and to many of the auxiliary assumptions.

To summarise the argument: multi-model ensembles are not prepared systematically or independently; they are ensembles of opportunity (W. S. Parker, 2013b, 2018; Tebaldi & Knutti, 2007). That is, they are constructed from whatever models existing research groups contribute, so

[27] In addition to the constraints of computational power, the size of these ensembles is constrained by human resources. Running simulations requires computer time, but building them requires people time, and climate modellers with adequate funding are a limited resource. This is a point to which I will return in Chapter 6.

agreement between them is not good evidence of a robust result. Of course, there should be some amount of shared structure in a family of models in order to satisfy the criterion of having a common causal core that remains stable across the family of models. In the present case, however, no close attention is paid to which structures should be included in the causal core and which parts should be varied, making it difficult to assess the extent to which ensemble studies strike the right balance between preserving and varying structures across models (W. S. Parker, 2011).

In response to these criticisms, modellers Fulvio Mazzocchi and Antonella Pasini (Mazzocchi & Pasini, 2017; Pasini & Mazzocchi, 2015) have argued that robustness analysis can be performed within climate science if modellers also consider alternative kinds of models (see also Katzav & Parker, 2015). Mazzocchi and Pasini describe two kinds of data-driven modelling frameworks that could provide independent lines of model-based evidence in climate attribution studies: neural network models (Pasini, Lorè, & Ameli, 2006; Schönwiese, Walter, & Brinckmann, 2010; Verdes, 2007), and Granger causality models (Attanasio, Pasini, & Triacca, 2012, 2013; Pasini, Triacca, & Attanasio, 2012). These models have structures that are very different from dynamic models like ESMs. Rather than producing results by solving equations that represent the physical processes thought to occur in the target system, neural network and Granger causality models produce their results by analysing relationships that hold within empirical data sets describing the target.

Data-driven models like the two kinds described by Mazzocchi and Pasini certainly add value to climate science. Their capacity to identify relationships in data sets that might have otherwise gone unnoticed allows them to provide extra support for claims like "20th Century warming would not have occurred without human activity," via a completely different representational route than the one taken by dynamic models like ESMs. However, Mazzocchi and Pasini's proposal does not offer a complete solution. For one, these data-driven models are just that: driven by data. Consequently, while they can be used in attribution studies—studies where the aim is to find the cause or causes responsible for some observed effect—they are not used for projection or forecasting, which are future-directed. ESMs, on the other hand, are often used for these purposes, but Mazzocchi's and Pasini's data-driven models would not be able to contribute to a multi-model ESM ensemble engaged in projection.

Moreover, and returning to an earlier point, robustness analysis requires that the members of the family of models contain a fixed causal core. To the extent that ESMs do share features, such as the dynamical core approximating the governing equations, these shared features would be lost when adding data-driven models into the ensemble.

In response to the dissimilarity between Levins' robustness analysis and ensemble studies, philosophers of climate science have suggested alternative conceptions of robustness, which they then argue do apply to climate modelling, though in ways other than ensemble modelling. My own line of argument in sections 3.4 and 3.5 is a version of this approach. Before moving on to that, I will take a moment to describe the alternative robustness conceptions and applications already found within the literature and belonging to Lloyd (2010, 2015) and Winsberg (2018b, 2018a).

Lloyd (2010, 2015) proposes the notion of *model robustness*, which is intended to emphasise the role of empirical evidence in robustness analysis. Lloyd cites Levins' (1993) response to Orzack and Sober in which Levins clarifies that model-building cannot have confirmatory power without the input of empirical data, arguing that empirical data guides modellers in making and justifying their choices of a causal core and auxiliary assumptions. If we have an ensemble of ESMs, Lloyd argues, these ESMs are themselves composed of many smaller computational modules, such as modules representing atmospheric or oceanic circulation and parameterisation schemes representing cloud formation or sea ice. While some of these modules may be shared between groups or have a shared history, many of them will not be. Moreover, the different decisions made by each modelling group when building the modules that compose their ESM may require the use of different data during calibration. These means that the members of an ensemble each have been tested against available data before inclusion in the ensemble and achieve an epistemically significant degree of independence as they have often been tested against different data sets pertaining to different measurable aspects of the climate system.

This argument can be combined with one Parker makes about the evidence that computer simulations produce to bolster the view that agreement among multi-model ensembles should increase

our confidence in supported hypotheses. On the question of what kind of evidence computer simulations produce, Parker (2020a) argues that simulations do not provide direct evidence in favour or against tested hypotheses, but that they can instead provide higher-order evidence. This is because, if the construction of a computer simulation model involves the use of empirical data, and a simulation produces a result supporting some hypothesis, then this suggests that the data supports that hypothesis. Of course, since the data is used in conjunction with various modelling assumptions used to construct the model and produce the result, the simulation really produces evidence that the data plus some assumptions support the hypothesis rather than the data alone. Here, again, we are at square one where we might want to know how reliable these assumptions are and, in an ideal world with infinite time and computational power, could perhaps vary these assumptions systematically, but instead must be suspicious of these assumptions. Nevertheless, Lloyd's argument does help with the difficulty if it's the case that many of these assumptions are themselves supported by empirical data and by different sources of data across different models.

While the worry of bad assumptions might still stop us from taking agreement among models in a multi-model ensemble as an indicator of accuracy, we might, using Lloyd's arguments, think that their agreement raises our confidence somewhat regarding their accuracy and their ability to demonstrate that, among our empirical data, we have evidence for the supported hypotheses revealed by our models. Robustness need not be an all or nothing thing but can instead come in degrees.

Model robustness is undeniably a desirable property and, I think, part of the explanation of why we should take the results of ESMs seriously. However, it is not the only way to apply robustness analysis to climate modelling nor the only way to do so that is particularly faithful to Levins' original message. First, while Levins did not explicitly advocate for the use of simple models in his (1966) paper, instead advocating for the use of general models, he nevertheless built and investigated simple models, belonging to a group called the "simple theorists" by E. O. Wilson (Chisholm, 1972, p. 177; c.f. Odenbaugh, 2006). Discussions about how well robustness analysis applies to ensembles of ESMs or individual ESMs and their component parts, therefore, depart from Levins' original intention to provide a methodology for building simpler models that, as he says in (1993), help us to understand

the system. This is, of course, not a problem. Levins' original intention does not matter all that much if some of the ideas can be taken in different directions for different purposes and value extracted in virtue of these departures. Nevertheless, I wish to steer the conversation in the philosophy of climate science in a different direction, away from more complex models and towards simpler ones because I think there is value to be found in doing so. Although there is nothing wrong with the direction discussions of robustness analysis have gone in the philosophy of climate science, fertile ground awaits in the other direction.

Before turning to my own view, I will present one final view of robustness in the philosophy of climate science, which I will actually make use of in my own account. Winsberg (2018b, 2018a) has recently advocated for Schupbach's (2016) notion of *explanatory robustness* as this as the right view of robustness for climate science. On Schupbach's view, robustness is fundamentally a matter of ruling out competing hypotheses. To illustrate, consider Ian Hacking's (1983, p. 193) discussion of early microscopes, produced in the 18th and 19th centuries, which were plagued of artefacts and distortions. Chromatic microscopes, for example, presented subjects as though they possessed vibrant and otherworldly colours. To overcome this problem, achromatic microscopes were made using flint glass lenses. Although the colourful distortions disappeared, microscopic subjects now appeared to be covered in little globules. It was long debated whether this observation was due to a real feature of the organic material under investigation or whether it was due to another distortion introduced by the instrument. New microscopic lenses later demonstrated that the globules were an artefact of the flint glass lenses. In this case, each new line of evidence tests a competing hypothesis: are the vibrant colours a consequence of properties of the subjects or the chromatic lenses, are the globules a consequence of the properties of the subjects or the flint glass lenses, and so on.

I agree with Winsberg's optimism regarding Schupbach's account. To a large degree, I do think this gets to the heart of robustness analysis and is commensurate with Levins' view of building model families. Each member of the family can be seen as testing a hypothesis: was the result caused by auxiliary assumption A_i? A modeller can then go down the list of family members and potentially problematic auxiliary assumptions. While I endorse Schupbach's notion of explanatory robustness, his

account is very general, applying to all kinds of enquiries and disciplines. This generality is a feature of the account rather than a bug, but I want to describe a narrower conception of robustness analysis. This is because I find the narrower conception that I will describe in section 3.4 to be more helpful with respect to the epistemology of modelling as it comes packaged with a method for testing the sorts of hypotheses that Schupbach and Winsberg describe. We will turn to that narrower conception of robustness analysis now.

3.4 Isolating causes with model spaces

In this section, I will present what I take to be an unexplored way of applying Levins' method of robustness analysis to climate modelling, at least in the context of the philosophy of climate modelling. First, let me begin with two senses of robustness analysis that Tarja Knuuttila and Andrea Loettgers (2011) argue can be drawn out of Levins' paper. The first sense Knuuttila and Loettgers call *independent determination* robustness and the second sense is called *causal isolation* robustness.

Independent determination robustness involves increasing researchers' confidence regarding a result by using multiple lines of evidence. In the present context, these multiple lines of evidence could take the form of multiple different models but in other contexts it could be generalised to include empirical sources of evidence such as the use of multiple experimental paradigms or different observational and measurement equipment (Wimsatt, 1981). Mazzocchi and Pisini's argument that ensembles would be improved by the inclusion of data-driven models in addition to the dynamic models that already comprise ensembles is directed toward this kind of robustness: including data-driven models would increase the different lines of evidence available to a researcher. Likewise, Parker's argument that multi-model ensembles are insufficiently independent is an argument against ensembles succeeding at this brand of robustness analysis: the shared history of models means that they cannot act as independent lines of evidence.

Causal isolation robustness, on the other hand, targets the causal mechanism driving the robust result rather than focusing on increasing a researcher's confidence in a result (Knuuttila & Loettgers,

2011, pp. 777–778). When performing causal isolation robustness analysis, multiple models are used as a means of investigating a possible mechanism or causal dependency, varying parameters or components to assess which combination of factors are sufficient to produce the focal result. As they see it, a key difference between this kind of robustness analysis and independent determination robustness, is that causal isolation robustness is often used when a result is already known, but the drivers of the result are yet to be determined. Contrast the use of multiple ESMs to forecast the likely consequences of different emissions scenarios, where the result of the scenarios is unknown, to a case where multiple models are used to investigate why some phenomenon occurs.

Although Knuuttila and Loettgers intend for this sense of robustness to extend beyond mathematical modelling and simulation, when restricted to the investigation of mathematical structures, the central idea behind causal isolation robustness is similar to some other views in the literature which likewise avoid considerations of confirmation (Forber, 2010; Odenbaugh & Alexandrova, 2011). Patrick Forber (2010), for example, denies that robustness analysis can increase our confidence in model results and instead characterises robustness analysis in the context of formal evolutionary theorising as an exercise of how-possibly modelling: "The formal inquiry exemplified by robustness analysis and simulation provides global how-possibly explanations that constrain what counts as a biological possibility" (p. 37). When applied only in the context of modelling, the sort of causal isolation that Knuuttila and Loettgers describe is also targeted toward uncovering possible dependence relations, focusing on those conditions that are necessary to produce the result and those under which the result will start to break down. The process cannot confirm an explanation of a known pattern or phenomenon but is used to develop and refine hypotheses which can then be tested empirically.

The causal isolation reading of Levins' robustness analysis, as opposed to the independent determination reading, has a clear connection to the notion of a robust theorem, described in section 3.2. Recall that a robust theorem was an articulation of a dependency that had the form "*Ceteris paribus*, if [common causal structure] obtains, then [robust property] will obtain" (Weisberg, 2006b, p. 738). We can think of the dependencies described by robust theorems as statements of the kinds of

causal relationships that are investigated through the process of causal isolation robustness analysis. Finding what conditions are sufficient to produce an effect, as one does in the context of Knuuttila and Loettgers' causal isolation robustness is tantamount to finding what common causal structures are sufficient to bring about the consequent of a robust theorem—that is, isolating a causal core—as well as finding details regarding the *ceteris paribus* conditions of the robust theorem and charting the scope and limits of the focal causal dependency. The result of causal isolation robustness, then, should be the kind of information used to construct robust theorems.

Let's bring the discussion back to climate modelling. At the end of section 3.3, I endorsed Schupbach's explanatory robustness as the right view of robustness analysis in general but noted that we could articulate a robustness concept with a narrower scope, which would be more directly applicable to the epistemology of models. Recall that Schupbach's explanatory robustness involves eliminating competing hypotheses, producing a hypothesis that is explanatorily robust if it has withstood this process. Causal isolation can be seen, at least in part, as an instance of this general process where the hypotheses being compared are ones regarding a focal causal dependency and the minimal conditions that are sufficient to produce the focal result. To demonstrate this, I will spend most of the remainder of this section presenting two instances where models are used to eliminate a competing hypothesis in such a way that isolates a causal dependency. Before moving on to those examples, however, let me also return to climate science's "model spaces," which I introduced at the end of Chapter 2. In what follows, I will show how such spaces can be used to guide model adjustment, test competing hypotheses, and isolate causal dependencies. This will further illustrate the value of reorienting the philosophical discussion of robustness analysis in climate science from the domain of ensemble modelling and toward an alternative modelling strategy.

Jeevanjee et al.'s (2017) multi-dimensional model space has six dimensions along which climate models can vary in complexity, which are shown in table 3.2. Note that this space is not intended to exhaust the ways in which processes or structures within a target can vary, or the processes and structures that ought to be varied. Looking at Jeevanjee et al.'s proposed space, we can see that, with respect to representations of the Earth's surface, for example, a model that represents real land

masses and sea ice is more realistic than a model that represents the Earth as an aquaplanet, completely covered with oceans. This is because an attempt has been made in the former case to include structures that make a causal contribution to the behaviour of the target system. However, there is no strict ordering among all the models that could possibly populate these spaces in terms of their complexity. While it is obvious that a model maximally complex along all dimensions is more complex than a model maximally simple along these dimensions, it is not obvious that a model maximally complex along one dimension and maximally simple across all the rest is more complex than a model maximally complex along some other dimension and maximally simple along the others. Likewise, if one model is maximally complex along three of Jeevanjee et al.'s six dimensions and maximally simple along the other three, and another model is moderately complex along all six, it is not obvious which is more complex than the other.

Fluid	**Rotation**	**Ocean**	**Surface**	**Convection**	**Radiation**
Compressible	Coriolis	Dynamical	Land + ice	Explicit moist	Spectral
Hydrostatic	β-plane	Column	Real land	Super-param.	Gray
QG	f-plane	Slab	Ideal land	Parameterized	Newtonian
Static	None	Non-uniform T_s	Aqua	Large-scale	Fixed
		Uniform T_s		dry	

Table 3.2 Jeevanjee et al. propose six dimensions along which climate models may be more or less complex and realistic.

Although I will focus on Jeevanjee et al.'s proposed space, it is worth noting that other spaces have been proposed. For example, Maher et al. (2019), who follow Jeevanjee et al.'s lead, propose a

four-dimensional model space for the atmospheric component of climate models, breaking the component down into the equations that represent fluid flow and three dimensions for the representation of physical processes. The mention of Maher et al.'s work is useful for demonstrating that, in complex models like ESMs, which are composed of many submodules coupled together, model spaces can be devised that focus on each of these modules. In the case of Maher et al.'s space, for example, the focal module is that representing atmospheric circulation, but there might be other ways that would be useful for different research aims. Decomposing a more complex model and model space into modules and their "regions" within the greater model space can be used as a method to continue representation of systematic variation of a model when it might otherwise become intractable.

How can a space like the one described in figure 3.2 be used to conduct causal isolation robustness? To answer this question, I will present Jeevanjee et al.'s (2017, p. 1764) demonstration that hypothesis testing is a function of the model space. This is relevant to causal isolation robustness because this is understood as a practice of producing hypotheses about causal dependencies, and the hypotheses that Jeevanjee et al. describe are precisely that. They give two examples, which I will report here.

The first example hypothesis concerns polar amplification. This is a well-known climate phenomenon whereby the effects of increased atmospheric CO_2 levels, such as increased average surface temperatures, are more exaggerated toward the poles than they are at lower latitudes. What explains this phenomenon? One possibility is the presence of an albedo feedback process initiated by highly reflective polar ice. The albedo would decrease as rising surface temperatures cause the ice to melt, allowing more solar radiation to warm the area even further, producing the exaggerated warming effect. A model study by Alexeev (2003) shows how such a hypothesis can be investigated by systematically varying a model and moving through model space. In their study, Alexeev build an aquaplanet model, which represents the planet without any land or ice, so represents the surface as having a uniform albedo rather than concentrations at the poles. In terms of Jeevanjee et al.'s six-dimensional space, this is a movement down the "surface" dimension to the point they label "aqua." In

this model, polar amplification is still observed but the typically-assumed caused has been removed. Consequently, the hypothesis that the distribution of reflective surfaces is the primary causal driver of polar amplification is brought into question and room is made for alternative hypotheses, which can be articulated and tested through further modelling (see Alexeev, Langen, & Bates, 2005).

Jeevanjee et al.'s second example of hypothesis testing with model spaces involves the phenomenon of Atlantic multidecadal oscillation. The Atlantic multidecadal oscillation is the cycle of variation in sea surface temperatures of the Northern Atlantic Ocean, which has a period of about 70 years. As with the first example, there is uncertainty regarding the causal driver responsible for the Atlantic multidecadal oscillation. One popular contender is that the phenomenon is caused by Atlantic meridional overturning circulation, which is the section of ocean circulation that takes place in the North Atlantic. Amy Clement and colleagues (2015) tested the hypothesis that Atlantic meridional overturning circulation was the primary driver of the oscillation by moving through model space along Jeevanjee et al.'s "ocean" dimension. In their study, general circulation models[28] in which atmospheric components coupled to oceanic circulation components were adjusted such that the atmospheric components were instead coupled to non-circulating, or "slab," ocean components. Clement et al. found the Atlantic multidecadal oscillation was stable across the two conditions. Consequently, the phenomenon could not be explained primarily by ocean circulation, even if heat transfer between atmosphere and ocean was still part of the explanation.

The two examples just presented illustrate how movement through a model space can be used to test hypotheses about causal relationships and dependencies. Let me now bring causal isolation robustness analysis back into the picture. Recall that, rather than being a method of confirmation or increasing confidence in a result via independent routes to the same result, causal isolation robustness is a method of building multiple models to investigate possible causal dependencies. Unlike in the case

[28] As mentioned in a previous chapter, general circulation models differ from ESMs in that ESMs represent more processes of the Earth's climate. Namely, ESMs represent the carbon cycle, including the land's and ocean's capacity to function as carbon sinks, while general circulation models do not.

of independent determination robustness, a shared result across multiple models is not used to increase confidence in the result. Rather, the differences between the models, including the processes they include and, importantly, omit, is used to provide insight into the minimal conditions that might explain the target pattern. By varying parts of a model, modellers can determine which factors suffice to produce a result. In the first example described above, Alexeev varies the representation of the planet's surface to demonstrate that uniform albedo suffices to produce polar amplification, so the uneven distribution that sees the poles being much more reflective the rest of the planet's surface is not necessary to produce the phenomenon. Likewise, coupling an atmospheric circulation component to a slab ocean rather than a dynamic ocean as in the Clement et al. study demonstrates that Atlantic meridional overturning circulation is not required to produce the Atlantic multidecadal oscillation.

Within the philosophy of climate science, there has been a significant amount of debate about how robustness analysis applies to climate modelling and whether it can help explain how the results of models are justified. In this section, I have argued that an alternative way of applying Levins' robustness analysis to climate modelling is to those studies involving more idealised models or a combination or more realistic and more idealised models. Climate science's "model spaces" allow for the systematic exploration of model space and, in doing so, facilitate the investigation of causal dependencies and minimal conditions that are required to reproduce phenomena.

Before concluding this chapter, I wish to consider the notion of scientific understanding, connecting it to the discussion so far regarding robustness analysis and model spaces. Recall from the last chapter that the comprehensiveness of brute force climate models came at the cost of understanding, with climate scientists themselves noting that simpler or more idealised models were often more useful in promoting understanding. When presenting Levins' robustness analysis in section 3.2, we also saw Levins' remark that a key aim of robustness analysis was to increase understanding of a system. Consequently, it would be valuable to end this chapter with a more precise conception of scientific understanding and an explanation of how this conception relates to both exploring model spaces and to causal isolation robustness. My account of the relationship between understanding and

causal isolation robustness will be somewhat of a sketch but lays the groundwork for future work that would more thoroughly explore and articulate these concepts and their relationship.

3.5 Scientific understanding

Scientific understanding is an increasingly popular topic among philosophers of science (see also De Regt, 2009; Dellsén, 2016; Elgin, 2017; S. R. Grimm, 2011; Potochnik, 2017), and, as I mentioned, scientists themselves speak favourably of understanding (Charney, 1963; Jeevanjee et al., 2017). The previous chapter espoused the view that complex and complicated computer simulation models do a far poorer job of facilitating understanding than simpler models, which was an epistemic weakness of such models and a reason not to pursue the strategy of complex modelling alone. Finally, as mentioned in section 3.2, Levins claimed that understanding could be achieved by performing robustness analysis. It is worthwhile, then, to spend some time considering just what scientific understanding might be and to explore the connection between understanding and robustness analysis. Despite extensive literatures on both topics, the connection between the two has received relatively little attention. The aims of this section, then, are to (1) characterise understanding, (2) demonstrate why fostering understanding is a goal of modelling, and (3) clarify the connection between understanding and robustness analysis.

3.5.1 Characterising understanding

As you would expect, there are many views of understanding in the philosophical literature. These views can be divided into factive and non-factive views. Roughly, factive views see understanding as an epistemic aim which succeeds in virtue of getting at the truth, but which involves aspects beyond truth simpliciter. For example, it might amount to being able to provide true answers to relevant questions (Khalifa, 2020), or using highly idealised models that are interpreted such that they produce true beliefs in the users (Rice, 2016, 2019; Ross, 2021). In contrast, non-factive views place the emphasis on the non-truth aspects and see understanding as an epistemic aim which is not primarily

directed toward reaching the truth. For example, understanding might be a skill or know-how rather than a kind of know-that (De Regt, 2009; Elgin, 2017), in which case it would be achieved by acquiring an ability rather than a set of true beliefs.

My own view sits within the family of factive views and sees understanding as the knowledge of an explanation (W. S. Parker, 2014; Strevens, 2008, 2013) or the grasping of a causal dependency (S. R. Grimm, 2011, 2014; Potochnik, 2017, p. 116).[29] To put this in plain language, there is a thing in the world, such as the target pattern or system to be understood and a representation that enables the scientist to manipulate or make interventions on that dependency to produce a result that they expect.[30] Here is how Stephen Grimm describes understanding and grasping (S. R. Grimm, 2011, p. 89; c.f. Potochnik, 2017, p. 114): "to grasp how the different aspects of a system depend upon one another is to be able to anticipate how changes in one part of the system will lead (or fail to lead) to changes in another part" (see also Dellsén, 2018). The COVID-19 conspiracy theorist, for example, does not understand the pandemic. They have a rich mental model about the causal relationships that are relevant to the pandemic but have only misunderstanding because, in reality, there is no pattern causally connecting Bill Gates and 5G towers to the disease. Consequently, if the conspiracy theorist were to remove Bill Gates and 5G towers from the causal network, the pandemic would, counter to their expectations, be no less severe.

That's the basic view. Now, let's get into some of the details.

When I speak of a causal dependency, I have in mind the conditions of Woodward's manipulability approach to causation (Woodward, 2003). This view dispenses with the metaphysical

[29] As I see it, "grasping a causal dependency" sits within the family of explanation-based views of understanding because causal dependencies are just the sorts of things that carry explanatory information, though some philosophers disagree that modelling dependencies amounts to possessing an explanation (Dellsén, 2018).

[30] An important caveat is that the manipulations or interventions may be only hypothetical as some interventions will be impossible, such as interventions on systems or processes very far away in space or time. More on this caveat below.

commitments about causation, focusing instead on how causation appears to be viewed by scientists from a range of disciplines, allowing us to speak of causation occurring in higher-level systems, like economic systems, without concern. The basic idea of Woodward's (2003) conception of causation is that *X* is a cause of *Y* if manipulating *X* can be used as a means to reliably manipulate *Y*. The presence of 5G towers does not cause the spread of viruses as intervening on the presence of 5G towers is not a reliable means by which to alter the presence of viruses. Greenhouse gas emissions cause an increase in global average surface temperatures because manipulating the presence and concentration of greenhouse gases in the atmosphere is a reliable way to manipulate global average surface temperatures. *X* and *Y* need not be binary variables. Consider the causal connection between emitting greenhouse gases and increasing global temperatures: you emit more or less and the resulting change in temperatures can be greater or smaller. Likewise, *Y* can be put in terms of a probability distribution: the greater the emissions, the higher the probability of severe weather events.

Of course, an agent need not *actually* be able to make interventions on a system whose relevant causal structure they understand (Woodward, 2003, pp. 10–11). For instance, I might understand the climate system to the extent that I recognise that reducing greenhouse gas emissions to zero and using negative emissions technologies to pull CO_2 out of the atmosphere would result in a lower global average surface temperature than if I let emissions follow a business-as-usual course, but I have no power to actually limit the Earth's emissions myself. Consequently, the claim must be understood counterfactually: if someone were to intervene on emissions as described, the climate would respond thus and so; if someone were to eliminate Gates and 5G towers, the pandemic would be no different.[31]

The next part of my view of understanding is the representations of these causal dependencies. To achieve scientific understanding, scientists need ways of representing relationships and

[31] An example of such a counterfactual in the historical case might be something like: had the Earth not been struck by an asteroid at what is now known as the end of the Cretaceous period, the evolutionary history of the planet would have gone in direction x.

manipulability conditions. Typically, they use models which exemplify these causal relationships. However, models contain many departures from reality, which are frequently intentional and thought to improve the models for different reasons (e.g. Wimsatt, 1987). To ensure that unrealistic models facilitate genuine understanding (as opposed to misunderstanding) they must be, as Angela Potochnik puts it, *epistemically acceptable*. For Potochnik, "A posit is epistemically acceptable when its divergence from truth is insignificant, taking into account (a) the posit's role in the representation and (b) the epistemic purpose to which that representation is put" (2017, p. 100). Take a 0D EBM, for example. The posit that Earth emits radiation as though it were a black body is a divergence from the truth. However, it is epistemically acceptable because it is insignificant if we want to either exemplify the key relationships that determine the planet's average temperatures or even make a relatively accurate estimation of changes in temperatures in response to changes within these key relationships, such as after increasing the albedo factor. On the other hand, the 0D EBM and its divergences from the truth would be epistemically *un*acceptable if we were concerned with the relationship between rising temperatures and the albedo effect—a hotter world means less snow and ice, so a lower albedo—and a more accurate estimation of the change in average temperatures than inclusions of such a feedback mechanism could permit.[32] Alternatively, consider the aquaplanet model in Alexeev's exploration of the polar amplification, described in the previous section. In this case, the posit that Earth's surface is a uniform ocean is an obvious departure from the truth, but it is one that is useful rather than harmful because the model is being used to investigate whether polar amplification persists in the absence of the uneven distribution of reflective surfaces.

Understanding a dependency does not simply amount to possessing a representation of that dependency that allows the possessor to accurately hypothesise about possible interventions on the dependency. Rather, it involves internalising such a representation, or developing a mental representation that the possessor can manipulate (Wilkenfeld, 2013). At that point, the scientist has "grasped" the dependency. This grasping or internalisation of a model is what I associate with the

[32] Supposing that this feedback mechanism was modelling accurately.

sense or feeling of understanding. However, as Jonathan Trout (2016) argues, a misplaced feeling of understanding can easily lead people astray. The conspiracy theorist may find themselves struck by the sense of understanding after digesting hours of videos about COVID-19: "Aha! *Now* it all makes sense: Bill Gates wants to control our minds with 5G towers!" But if the model internalised is not epistemically acceptable, then there is no genuine understanding.

The condition that understanding requires the possession of a mental model reveals why simpler models can be, all things being equal, better at facilitating understanding. While they are suited toward supporting understanding because they make the dependency and some of the main ways in which it can be manipulated salient, simpler models are also easier to internalise. A climate modeller could easily internalise a few energy-balance models, confidently stating, when asked, the key relationships that can be introduced to the models and writing the formal descriptions of such models down. As far as I am aware, no one could do such a thing with a high-fidelity Earth system model.

Note that mere internalisation is not enough if the possessor has no proficiency with the representation. That is, they cannot simply memorise some facts without knowing how the facts relate. If they have understanding, then they are able to manipulate the model successfully and alter the representation for nearby cases. For example, I know the equations specifying the Lotka-Volterra predator-prey model. I could write them down if you asked me to and I could tell you what the terms mean and why they have the relation to one another that they do and why manipulations of the model produce the results they do. I have some understanding. But I do not have the degree of understanding that would allow me to expand the model in the various ways that it has been expanded by theoretical ecologists. It is in this respect that the insights of non-factive views of understanding, some of which focus on skills and abilities are useful. The notion of grasping, as I see it, captures these intuitions about the skill a scientist might have with building, manipulating, and making inferences with a model to successfully represent a target relative to their aims.

So, that's scientific understanding. There is a pattern or dependency to be understood. There is a model that represents that dependency such that relevant manipulations of the model match relevant manipulations to the model. This allows the modeller to correctly anticipate the responses of the dependency in response to manipulations (including hypothetical ones). Finally, the modeller has a proficiency with the representation to the point where they have internalised the relationships of the model and manipulate the model mentally and know what to do with a slightly different model or something. It would, of course, be possible for someone to achieve scientific understanding in the absence of a formal model, instead constructing a mental model with the relevant properties directly. External models, specified formally, are often preferred simply because of their ability to rigorously encode and embody intervention relationships.

Next question: why is understanding something valuable?

3.5.2 Why understanding?

Hopefully, the characterisation of understanding provided gives some clues as to why scientific understanding is valuable. Indeed, if one has a factive view of understanding as I do, then it is easy to motivate the utility of understanding: getting at the truth is generally thought to be a scientific aim, and understanding is just relevant truth that allows you to (excepting the earlier caveats) manipulate the world. That said, understanding is a useful concept precisely because it accommodates valuable falsity in descriptions of the world. One reason to be attracted to understanding as an epistemic goal of science is that it allows us to explain why the accomplishments of Isaac Newton's physics remain an epistemic achievement despite the theory being superseded by Albert Einstein's physics, which will itself be superseded in the future (Dellsén, 2016). That is, these theories offer descriptions that increase our ability to manipulate the world successfully to produce results that we expect. For example, Newton's work was used in calculating flight trajectories in the Apollo lunar missions (Bennett, 1970).

Throughout the remainder of this subsection, I will consider the value of understanding in the specific context of complex simulation modelling. So far, I have taken the view that complex models are not well suited to promoting understanding. Indeed, they can obscure understanding with their complexity. This view is shared by Gabriele Gramelsberger and colleagues (2020), who argue that understanding becomes more difficult when dealing with high-fidelity climate models because of their complexity. As I see, understanding is difficult to achieve with highly complex models for at least two reasons. First, the more complex a model is, the less it can isolate the key causal dependencies that are central to understanding. And, second, the more complex the relationships included in a model, the more difficult it is for a scientist to grasp or internalise these relationships.

Although highly complex models are not apt for supporting understanding, understanding their targets—perhaps through other models—remains essential. This is because understanding the target of a highly complex model can assist both in the construction of the model and, crucially, its evaluation and interpretation. As Volker Grimm puts it: (1996, Chapter 152): "…prediction without understanding represents blind faith in the power of computers."[33] While understanding dependencies might not be sufficient to allow a modeller to make precise predictions, they should provide some insight into qualitative changes that put constraints on what kind of results appear consistent with the existing picture of the system and which do not. This is not limited to the model as a whole but can also apply to evaluations of sub-components within the model. Remaining on the topic of model

[33] I have omitted the first part of this quotation where Grimm et al. state that "understanding without the ability to predict is illusion." There, Grimm and colleagues are directing their criticism towards theoretical work in ecology. As the debate between Levins and the systems ecologists described in section 2.1 indicated, the applicability of theoretical ecology has long been a concern in the field. According to my characterisation of understanding, Grimm et al. are correct in so far as correctly anticipating responses to hypothetical interventions is a key part of understanding. If there is a sufficiently large disconnect between theoretical models and target phenomena, then it's entirely possible for understanding of a model to create the illusion that the phenomena has been understood, but there will be no genuine understanding if manipulations of the model provide little to no information about counterpart manipulations of the target system.

evaluation, complex simulation models are themselves objects of understanding, which require investigation in order to build, improve, and use properly. According to Gramelsberger et al. (2020), modellers describe acquiring a "feeling" for what sorts of adjustments might produce what results in an ESM even if it's not the sort of understanding that can be acquired from simple models. They call this kind of acquaintance "pragmatic understanding." So, even if high fidelity models are not especially useful for fostering understanding, they must be understood by modellers to some degree in order for modellers to get the most out of them.

That said, it may be wrong to conclude that highly complex models contribute nothing to understanding. As Parker (2014) suggests, computer simulation has improved understanding because it allows people to test hypotheses about causal dependencies and necessary and sufficient causes. While it's possible that this is only possible because there is more foundational knowledge supported by simpler models, it is undeniable that complex models provide a testing ground for hypothetical interventions and hence for exploring causal dependencies.

So, understanding is a useful concept as it captures the value of partial truth and idealisations in scientific representation and understanding itself is a useful epistemic achievement as it is knowledge that can be applied to the world to make interventions. Understanding a phenomenon with simple models is useful for building and evaluating complex models, which, due to their capacity to represent hypothetical interventions, may contribute somewhat to understanding even if it is not their primary purpose. Now, let's turn to the relationship between understanding and robustness analysis.

3.5.3 Understanding and robustness analysis

With understanding and its value described, let's turn to the final question of the connection between understanding and robustness analysis. Continuing the line of argument from section 3.4, I will be restricting my discussion to the connection between understanding and causal isolation robustness. Recall that causal isolation robustness involves investigating a causal dependency and ascertaining what conditions suffice to produce a focal effect. This is a method for refining hypotheses about the

causal drivers of observed patterns, eliminating those positing unnecessary conditions and identifying more plausible causal dependencies which can then be tested empirically. As argued in section 3.4, model spaces track the ways in which target structures and processes can be represented within a model. These can assist with causal isolation robustness by providing a framework for exploring which conditions are necessary, unnecessary, sufficient, and insufficient, as well as the range of conditions over which the focal dependency holds. In effect, exploring these conditions helps scientists to anticipate how a dependency will react in the face of different interventions.

Causal isolation robustness contributes to scientific understanding as I have characterised it in this section in at least two ways. First, identifying causal dependencies is central to scientific understanding and is also the aim of causal isolation robustness. Although causal isolation robustness alone cannot result in understanding as, without empirical confirmation, causal isolation robustness cannot sort plausible from actual causal drivers, it is, nevertheless, well suited to articulating plausible dependencies, which can then be tested. The exploration of model spaces assists further in demonstrating the limitations of these plausible dependencies, illustrating to the modeller in just what conditions a dependency holds and how these causal relationships react to interventions on other aspects of the system. So far, this is not a unique view of robustness analysis and has much in common with Forber's view that robustness analysis is an exploratory exercise that can be used to develop and refine how-possibly explanations. The difference here is the further step that recognises the contribution of possible and plausible causal explanations to the achievement of scientific understanding.

If the first way in which causal isolation robustness contributes to understanding is through the identification of plausible causal explanations and dependencies, then the second way is through assisting scientists with the task of grasping these dependencies. Grasping causal explanations, on the view described above, is achieved when one is able to construct a mental representation of the dependency, often by internalising the relationships exemplified by a public representation like a scientific model, such that one is able to manipulate the mental representation proficiently (Wilkenfeld, 2013). To put it another way, systematically varying and exploring a model or family of

models provides a great deal of information about how those models behave, increasing modellers' skill with the model. Recall that some of the non-factive view of scientific understanding see understanding as the acquisition of a skill or know-how rather than know-that. Although I do not subscribe to these views, I believe they are right in emphasising the importance of the skilful use of representations scientists deploy to anticipate the behaviour of their targets. Causal isolation robustness, then, contributes to a modeller's skill with their scientific tools—that is, their models.

This discussion of understanding and robustness analysis has only been relatively brief. It is my belief that there is much more to say about the relationship between robustness analysis, which has long been discussed in the philosophy of science, and scientific understanding, which is an increasingly popular topic within the discipline. As I've stated, a more thorough examination would pursue two aspects: (1) robustness analysis as a method of articulating plausible causal explanations, which are central to many factive views of understanding in the literature; and (2) robustness analysis as a method which improves modeller's skilful use of their models, which captures some of the intuitions motivating non-factive views of understanding. For now, I must leave this topic and conclude my discussion of robustness analysis and climate modelling.

3.6 Conclusion

It's time to bring this chapter to a close. My primary aim in this chapter was to continue my investigation of Levins' (1966) work in order to find lessons for the epistemology of models and computer simulation. I examined the concept of robustness analysis, which has been discussed at length within the philosophy of climate science as well as the philosophy of science more broadly. My secondary aim was to present a view of scientific understanding, which I take to be a natural bedfellow of the causal isolation conception of robustness analysis.

In pursuing these aims, this chapter contributes to the philosophy of climate science. As section 3.3 demonstrated, there has been much debate regarding the connection between Levins' robustness analysis and climate modelling within this literature. My contribution has been to bring the

work of Knuuttila and Loettgers, themselves primarily philosophers of biology and economics, into contact with the literature in the philosophy of climate science. By using their concept of causal isolation robustness, I showed how the model spaces proposed by climate modellers can be used to investigate causal dependencies and identify causal relationships in accordance with causal isolation robustness. Both causal isolation robustness and model spaces also have a natural connection to the Levins' robust theorems, which, according to Weisberg, are *ceteris paribus* statements of causal dependencies. As I've argued, model spaces serve as frameworks that modellers can use to chart the ways in which their models can be varied and act as a guide for explorations of the conditions which are sufficient to produce a focal effect, assisting in the identification of a causal core and *ceteris paribus* conditions of a robust theorem.

Not only is this a contribution to the philosophy of climate science but to the philosophy of science more generally, which has also discussed robustness analysis at length. Within this literature, across the philosophies of biology, economics, and climate science, there has been a large amount of debate regarding the purpose and value of robustness analysis. Some have argued that it provides a kind of confirmatory power, while others have argued it has a more exploratory aim. The analysis I have undertaken, which uses climate science as a case study, has led me to conclude that one valuable way to think about Levins' robustness analysis—at least in the context of modelling—is the exploratory form directed toward the search for possible causal dependencies. I have demonstrated how model spaces can be used as a map of some of the landscape to be explored and that the systematic variation of models can be used to evaluate hypotheses about causal dependencies and the conditions sufficient to produce certain effects.

Moreover, I have suggested a connection between the claims of both Levins and proponents of climate science's model spaces regarding understanding. Adopting a view according to which scientific understanding is achieved when someone grasps a causal dependency, I demonstrated how systematically varying models with a model space would provide insight into a causal dependency and the effects of interventions, indicating which factors belongs in a causal core of a robust theorem and the *ceteris paribus* conditions of the focal dependency. Moreover, systematically varying models with

the assistance of a model space provides insight into the representation and its response to interventions, helping the scientist to grasp both the models and the dependencies they exemplify.

With respect to the epistemology of computer simulation, this chapter has laid some important groundwork on both robustness analysis and understanding that will inform Chapters 5 and 6. In the next three chapters, I will take a closer look at agent-based models, a class of computer simulation models, assessing the kinds of dependencies they can assist us in understanding as well as the epistemic challenges their structures create for understanding. As we will see, agent-based models sometimes struggle to foster understanding in part because a lack of standardised model construction and communication practices leaves many model structures less accessible than they otherwise could be. A consequence of this opacity is that it impedes model replication studies and robustness analysis, which further hinders their ability to support scientific understanding.

In the next chapter, however, I will look at a potential advantage of agent-based models, which is that they explain population phenomena mechanistically.

4

Agent-based modelling and explanation

In this chapter, I examine the representational capacities of agent-based models and investigate the kinds of explanations they support. Agent-based models are described by modellers as supporting mechanistic explanations. However, agent-based models and the systems they represent are typically populations, and populations are usually thought to resist mechanistic analysis. Instead, it is thought that they should be approached with "population thinking." This presents a puzzle. In this chapter, I explore the relationship between mechanistic explanation and population thinking by describing three ways in which population phenomena can be represented and explained. The third way, involving the use of agent-based models, has distinctly mechanistic features, including the decomposition of systems into their lower-level entities and a focus on the activities of those entities. However, there are also important departures from the mechanistic framework. My analysis results in a partial vindication of claims regarding the mechanistic capacities of agent-based models and suggests more work should be done on the relationship between population thinking and explanation.

4.1 Introduction

The previous two chapters have focused on models representing the Earth's climate, which operate primarily by approximating dynamical equations representing fluid flow. Examining these models raised epistemological questions about the costs of increasingly realistic representations and about how modellers can increase their understanding of processes by building and investigating families of systematically varied models. In the next two chapters, we will temporarily turn away from climate

models to examine a class of models with a very different kind of structure: agent-based models (ABMs).

ABMs are comprised of virtual populations of interacting units operating according to a set of sometimes quite simple rules. From their individual actions and their collective interaction, ABMs can produce quite surprising and complex behaviour at the whole-system level. [34] In this chapter, I will take a closer look at the representational and explanatory capacities of ABMs. As I will argue, there is a clear tension between the kinds of explanatory and representational capacities ABMs are sometimes said to have by practitioners and some philosophers, and the conditions of those explanations. Namely, many describe ABMs as *mechanistic*. The mechanistic approach has been popular in the philosophy of science over the past two decades and emphasises the importance of describing systems of interacting parts, focusing on the active, spatial, and temporal organisation of these parts, in order to explain and understand target phenomena. However, as representations of populations of interacting units, ABMs do not appear to fit within the scope of mechanistic explanation, which is better suited to tightly integrated systems where components are functionally differentiated.

At least at first glance, the decompositional aspect of the mechanistic framework fits ABM well, with many agent-based modellers proposing that a focus on entities and interactions, as opposed to population-wide variables and relations, is a central feature of ABM. Agent-based modellers José Galan and colleagues, for example, claim that the "defining characteristic" of ABM is that "entities within the target system to be modelled – and the interactions between them – are explicitly and individually represented in the model," and that this "is in contrast to other models where some entities are represented via average properties or via single representative agents" (2009, para. 2.7). Presenting their step-by-step guide to building ABMs in social psychology, Eliot Smith and Frederica Conrey express a similar view, recommending that modellers begin by thinking "theoretically in terms

[34] Agent-based models are known by different names in different disciplines. For example, ecologists tend to refer to these models as "individual-based models," while the term "agent-based models" is typically favoured in the social sciences. I will simply use the term "agent-based model" to refer to all these models.

of entities and interactions, not in terms of variables" (2007, p. 99). In a recent review of epidemiological models, Stephen Eubank and colleagues describe agent-based models as being able to "provide a high-resolution, mechanistic explanation of the reproductive number" (2020, p. 6).

Of course, we philosophers should be cautious about reading too much into the terminology of practitioners, who are unlikely to be aware of the peculiarities of terminology as it is used in the philosophy department. However, practitioners are not alone in describing ABMs as mechanistic. Long-time mechanist, Stuart Glennan, states that ABMs are "mechanistic models because these models show how a collection of individuals gives rise to some phenomenon that is the behaviour of the systems as a whole" (2017, p. 102). And he does not appear to be mistaken. If describing a mechanism for a phenomenon amounts to describing the parts, actions, and interactions underlying a target phenomenon, then ABMs appear to satisfy these criteria. Agent-based modellers typically have a target pattern in mind, such as neighbourhood segregation, the flocking of birds or schooling of fish, the spread of fire through a forest, the spread of a disease, and so on. They then describe a set of entities, such as households, birds or fish, trees, people, and so on. These entities have behavioural rules they follow, which dictate their activities and interactions with other entities. ABM also provides descriptions of populations that explicitly locate entities and their interactions within space, situating these entities within a representation of space, such as a lattice. This is particularly important given the centrality of the spatial organisation of components within the mechanistic explanatory framework.

So, prima facie, there is a reason to think that ABMs provide mechanistic explanations. Not only do modellers themselves take ABMs to provide such explanations, but they seem to fit within the philosopher's understanding of this terminology: mechanistic explanations focus on descriptions of systems of interacting parts, which are spatially located, and ABMs describe interacting components that are often located within a virtual space. However, populations—virtual or otherwise—have organisational features that are very much unlike those of systems typically thought to be amenable to a mechanistic approach to explanation (Levy, 2014; Levy & Bechtel, 2016; Wimsatt, 1986). Most notably, populations are not highly functionally integrated systems and frequently do not have rigid and stable spatial arrangements.

In this chapter, my aim is to examine the extent to which ABMs can be said to be mechanistic. All ABMs are decompositional, and are similarly focused on components, activities and interactions. However, many ABMs do not situate agents within space, and so this aspect of the mechanistic framework is missing. When ABMs do situate agents within space, and particularly when they situate them within a heterogeneous and relatively realistic virtual space, then they truly are mechanistic descriptions, despite being representations of populations and population phenomena.

My argument will proceed as follows. First, in section 4.2, I will provide further details on the mechanistic explanatory framework. In section 4.3, I will outline the problem with applying the mechanistic approach to populations. This is that the mechanistic approach has greater explanatory power for systems with spatial and temporal organisations unlike populations. In section 4.4, I will consider one possible solution suggested within the literature on populations and mechanistic explanation. According this response, mechanistic explanation can be applied to population phenomena via a model which represents those populations mechanistically. I reject this solution on the grounds that the models described lack the representational capacities to represent mechanistically. In section 4.5, I provide my solution, which is that the epistemic value of the mechanistic approach is not settled by the extent to which the system is machine-like. Rather, structural and organisational specifics of population phenomena can be important in many kinds of explanatory projects. I use two examples to show how ABM can be used to satisfy the epistemic aims and descriptive characteristics of mechanistic explanation. In section 4.6, I further demonstrate that different explanatory approaches can be taken when dealing with populations by describing three such approaches. I show that spatially implemented ABMs share the descriptive characteristics of mechanistic explanations and are unique among the modelling approaches taken toward populations in doing so.

4.2 The new mechanistic philosophy

The mechanistic account of explanation has been prominent in the philosophy of science for around two decades.[35] Views of explanation previously popular among philosophers, like the law-based deductive-nomological view (Hempel, 1965; Hempel & Oppenheim, 1948), did not appear to accurately describe scientific practices in some areas of the special sciences, such as molecular biology and neuroscience (Bechtel & Richardson, 2010; Craver & Darden, 2013). In the life sciences, for example, the presence of evolved complexity (Mitchell, 2009) or causal heterogeneity (Elliott-Graves, 2018; Matthewson, 2011) excludes meaningful universals. Rather than making scientific progress by unifying disparate phenomena under more fundamental laws, as might seem plausible for sciences like physics,[36] molecular biologists and neuroscientists instead make progress by, according to the new mechanistic philosophy, identifying systems of interacting components that are responsible for producing some observable effect (Glennan, 1996; Machamer, Darden, & Craver, 2000).[37]

Before we dive into the details of the mechanistic framework, let's start with a quick example of a mechanistic explanation. The Venus flytrap (*Dionaea muscipula*) is the most well-known of a small handful of carnivorous plant. How does it catch its prey?

To trap its prey, the plant must first lure prey to its traps, which it does through two signals. The first signal is colour-based, with the trap leaves turning red when the plant is hungry. The second signal is olfactory, with traps emitting over 60 volatile organic compounds, which are used by plants more generally to attract pollinators (Kreuzwieser et al., 2014). Many of these compounds are common constituents of fruit and flower scents and are particularly good at attracting hungry flies.

[35] Some book-length treatment of mechanisms includes (e.g. Bechtel & Richardson, 2010; Craver, 2007; Craver & Darden, 2013; Glennan, 2017).

[36] As Cartwright (1983) argues, unifying disparate phenomena with fundamental laws has limited explanatory power even in physics given the idealised nature of these laws.

[37] Note that the mechanistic approach is not the only alternative to deductive-nomological or law-based views of explanation.

Once a victim lands on the trap, it begins to search for its reward. In doing so, the insect touches trigger hairs, which are connected to sensory cells at their base and which convert physical stimulation into an electrical signal that serves as an action potential. Two action potentials within about 30 seconds are sufficient to trigger trap closure, and continued action potentials cause the trap to close further, ending with the formation of a stomach (Hedrich & Neher, 2018).

The rapid closing of the trap relies on both the geometry of the trap as well as cellular changes (Sachse et al., 2020). With respect to the geometry, the two lobes of the trap have two bi-stable states, one concave (their open state) and one convex (their closed state), which can change rapidly due to snap-through buckling. For an analogy, think of an open umbrella, which can quickly snap to an inside-out position when a sufficient force is applied to it, such as a strong wind. The cellular changes work in concert with the trap geometry and assist in triggering the snap-through buckling. This is an expansion of the lobes' outer epidermis and counterpart shrinkage of the inner epidermis. For this action to occur, it is important that the lobes are in a prestressed, "ready-to-snap" state, caused by "internal hydraulic pressure differences between layers" (Sachse et al., 2020, p. 16040). Venus flytraps in a dehydrated state do not regularly produce the rapid closure due to lacking the hydrological resources required to set the trap in its ready-to-snap state.

With this brief description of an example mechanism in hand, let's turn to some the philosophers that have described mechanisms and mechanistic explanation. Glennan characterises a minimal mechanism for a phenomenon as consisting of the "entities (or parts) whose activities and interactions are organised so as to be responsible for the phenomenon" (2017, p. 17). As Glennan admits, this minimal conception of a mechanism is very broad, and mechanisms in this sense are widespread. However, this sense of mechanism is shared by Carl Craver, who describes one of his examples as follows (2007, p. 5): "This is a mechanism in the sense that it is a set of entities and activities organised such that they exhibit the phenomenon to be explained." Returning to our example, the trapping mechanism described above is partly constituted by a set of mechanosensory hairs (entities) which interact with the prey as well as with the cells (more entities) at their base, engaging in the activity of sensing and signalling the presence of prey. Other entities that form part of

the trapping mechanism are the volatile organic compounds used to lure prey, the trap leaves themselves with their particular geometry, and the hair-like cilia that assist in preventing escape before the trap is completely sealed, and calcium ions (not mentioned in the description above) that assist in tracking the number of action potentials (Hedrich & Neher, 2018).

Further features beyond the minimal conception are, first, that mechanisms are often hierarchically organised, so mechanisms can be nested within other mechanisms and one mechanism can be a part, or component, of another. The trap leaves, for example, can be decomposed into their shell-like lobes, the mechanosensory hairs, the cilia around their edges, and the lobes can be further decomposed into the cells of which they are made, with the activities of layers of these cells responsible for the changing shape of the lobes from open to closed and back to open again. Second, "Entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration" (Machamer et al., 2000, p. 3). This is not simply to say that entities and activities exist in physical time and space. Rather, their spatiotemporal features are often crucial to the phenomenon to be explained and, if parts were shifted from their appropriate locations, the phenomenon would not occur. It would do the Venus flytrap no good if its mechanosensory hairs were anywhere other than inside the trap where it can signal the presence of prey in the trap. Likewise, the structure of the trap lobes matters for their rapid closing, and they must close rapidly to prevent prey escape. Finally, the activities and interactions of the parts are regular (Andersen, 2011, 2012; Machamer et al., 2000): the mechanosensory hairs can be relied upon to cause an action potential when stimulated.

Mechanisms, in summary, are things in the world best characterised as systems of parts that produce some phenomenon. However, in this chapter, we are not only interested in mechanisms but in mechanistic explanation. So, what exactly is a mechanistic explanation?

Mechanistic explanations are, mostly simply, explanations that cite information about the mechanisms that underly and produce a target phenomenon to be explained. According to Glennan, mechanistic explanations can be contrasted with "bare causal explanations" or "*what-but-not-how-*

explanations" (2017, p. 223). According to Glennan, "Bare causal explanations show what depends upon what without showing why or how this dependence obtains" (2017, p. 224). I prefer the latter term as "bare" suggests that they are missing something, which is not a perspective I wish to endorse even tacitly. Nevertheless, an example of a what-but-not-how explanation would be the following: *what caused 20th and 21st Century warming? Increased greenhouse gas emissions.* While this explanation is right, it does not tell us how increased greenhouse gas emissions produce increased global average temperatures. Rather, it simply states the dependency between the two.

This view of mechanistic explanations as *explanations how* creates two dimensions along which mechanistic explanations can be better or worse (Craver, 2007). First, the ideal mechanistic explanation of, say, the flytrap's closing, is a description of how it actually coordinates this behaviour, not a plausible description. That is, mechanistic explanations pursue how-actually rather than how-possibly descriptions. Second, complete mechanistic explanations are preferred to *mechanism sketches*, which leave parts of the process as black boxes, with a continuum of *mechanism schemata* lying between sketches and complete descriptions. The Venus flytrap example above, for instance, lies on this continuum as many details have been omitted, such as how the action potentials triggered by the mechanosensory hairs produce the cellular changes required to close the trap leaves. As explanations of how phenomena are produced, mechanistic explanations should aim for accurate and complete representation of the underlying system of interacting components. These two continua also indicate the development cycle of mechanistic explanations in science, starting with possibilities and sketches and ending with actualities and complete descriptions of the causal system.[38]

[38] Craver and Darden (2013) describe two further ways in which descriptions of mechanisms can vary. They can vary with respect to their abstraction and specificity and they can vary with respect to the breadth of their scope. However, unlike in the case of the degree of completeness or evidential support, they do not state that mechanistic descriptions ideally occupy one end of the spectrum. As Matthewson (2020) argues, and as Levins' arguments presented in Chapter 2 would suggest, degrees of abstraction and generality co-vary because omitting detail is a method for capturing a greater number of cases that would be excluded by the missing details.

Additional features of mechanistic explanation are that they are constitutive explanations which focus on parts, which are understood as physical entities. These physical entities are also organised. Craver (2007) describes three kinds of organisation. The first is active organisation, which charts which components interact with which other components and how. For example, the mechanosensory hairs produce action potentials which trigger changes in the cells comprising the trap leaves, which cause the trap leaves to close. The volatile organic compounds emitted by the flytrap do not interact with other parts of the plant but do attract prey, which land on the leaves. The second kind of organisation is spatial organisation: the mechanosensory hairs must be situated on top of the trap leaves so that their triggering can signal the current presence of prey within the trap, and the cells comprising the trap must be arranged such that they can take advantage of snap-through buckling to accelerate trap snapping. The importance of this kind of organisation likely explains the common accompaniment of visual representations to mechanistic explanations since they can be better suited than linguistic descriptions at representing spatial organisation (Bechtel & Abrahamsen, 2005). The third kind of organisation is temporal: activities must take place at a certain point in time, in a certain order, and for a certain duration. Stimulation of the mechanosensory hairs must take place before trap closing and two action potentials within about 30 seconds are required to trigger the flytrap as this makes for a reliable signal of the presence of prey.

Of course, one might think stating that mechanistic explanations focus on parts and organisation is tantamount to saying that they are descriptions of mechanisms, since these are comprised of parts and activities with the three kinds of organisation just listed. However, the focus on physical parts and these three kinds of organisation distinguish mechanistic explanations from other approaches to explanation which are also *explanations-how* and proceed by decomposing systems, such as the functional or computational explanations of cognitive psychology (Anderson, 2014; Cummins, 1975, 1983, 2000; Weiskopf, 2017). Unlike mechanistic explanations, however, these functional explanations decompose processes into sub-processes rather than decomposing physical structures and systems down into more physical structures and systems. You could decompose a cognitive process into its computational atoms without ever saying anything about the neurons that

perform those computations or even that there are independently identifiable physical parts that map onto these computational atoms. This can be a point a confusion as, within cognitive psychology, the word "mechanism" is used extensively to describe these functional and computational descriptions (e.g. Heyes, 2018a).

More broadly, the ideas of a "mechanism" and "explanations that describe mechanisms" are confused by a general practice of using the term "mechanism" to simply mean "causal driver." Returning to Glennan's distinction between what-but-not-how causal explanations and mechanistic explanations proper, one could easily consider any occupant of the "what" aspect of what-but-not-how explanations as a mechanism. *What caused 20th and 21st Century climate change? Increased greenhouse gas emissions. Increased greenhouse gas emission was the mechanism behind climate change.* In such a case, we have cited a causal driver but said little about underlying parts and their organisation. When, throughout the remainder of this chapter, we consider whether population phenomena are mechanisms amenable to mechanistic explanations, we must be cautious not to slip into speaking about mechanisms in this much broader sense of the term.

Let me be explicit about my own view of mechanistic explanation in relation to other kinds explanations. Above, I briefly compared mechanistic explanations with explanations that cite causal dependencies without decomposing systems at all (what Glennan called "what-but-not-how" explanations) and those that provide a decomposition according to functional capacities. When I make such comparisons, I do so only to be specific about the characteristics of mechanistic explanation. I do not wish to imply that mechanistic explanations are somehow superior to other forms of explanation. I take it that they will be superior given some purposes but also that other forms of explanation would be superior given some purposes too. That is, I endorse a pluralism toward kinds of explanatory frameworks.

Given this explanatory pluralism, you might wonder why it matters whether we can produce mechanistic explanations of population phenomena. If we take the position that mechanistic explanations are the epistemically superior explanations, then the motivation is obvious: bad news for

population phenomena if they can only be explained with inferior approaches to explanation. However, pluralism does not exclude the view that, for some purposes, what-but-not-how explanations suffice, for others functional explanations are preferred, and for others we require information about physical components and their organisation. One such case might be where we need to make an intervention on a system. It is one thing to know or to explain that we must reduce greenhouse gas emissions to reduce the extent of climate change, it is quite another thing to know or to explain how to make such an intervention. And doing so requires knowledge of parts: reduce cattle populations, focusing on removing them from arable land that could be used to farm vegetable crops; feed the remaining cattle on sustainably farmed seaweed but only if the cattle and seaweed are not so distant that transporting the feed to the cattle produces too much greenhouse gas emission; where possible replace aeroplane routes with low-emission, high speed rail; introduce a price on carbon so low-emission options appear more favourable to economic actors.

With the details of the mechanistic explanatory framework presented, I now wish to move on to the debate regarding populations and mechanistic explanation with ABM.

4.3 Populations and mechanisms

Numerous philosophers have criticised the scope of the mechanistic approach to explanation, arguing that it is apt only for systems that we can characterise as more "machine-like." Consider, for example, this quotation from Godfrey-Smith (2009a, p. 148) that contrasts "population thinking" with the representation of machines:

> When we embark on population thinking, we treat a system as an ensemble of individual things, which have some degree of autonomy and a significant number of properties in common. We should also know roughly where one ends and another begins. Some collections of things are too tightly integrated to be usefully seen as populations—the atoms in a hemoglobin molecule, for example. Other systems have parts that are too different from each other, and whose roles depend primarily on those differences—a car's engine. A highly

> structured network with heterogeneous and non-interchangeable parts is a different thing from a population. A riot, in contrast, is a populational phenomenon, as is the mixing of molecules in a gas.

If representing populations and engaging in population thinking is an exercise very different from engaging in what we might call "mechanism thinking," then this would spell trouble for the applicability of the mechanistic approach to population phenomena or the virtual populations of ABMs. In this section, I will present criticisms of the scope of mechanistic thinking and apply those criticisms to population phenomena and ABM.

Before we begin, though, I must make two caveats. The first is that, following Arnon Levy (2014), I will be using the term "machine-like" to describe systems or processes with a particular kind of organisation. This is to avoid describing some mechanisms—systems of parts acting and interacting to produce a phenomenon—as *more mechanistic* than others. Remember, mechanisms in the sense described in the last section are widespread. But, for those things that are systems of parts acting and interacting to produce phenomena, their active, spatial, and temporal arrangement can be more or less machine-like. The real question, then, is how valuable applying a mechanistic approach to explanation is for different systems. That is, how valuable is it to describe parts and their arrangement when a system is less machine-like? The basic answer to this question, outlined throughout this section, is that citing information about parts and their organisation is less valuable when systems are less machine-like because in those instances, information about organisational specifics matters less for the production of the explananda.

The second caveat is that mechanists like Craver (2007) have noted that "mechanism" should not be equated with "machine," recognising that the features of biological mechanisms depart in many important ways from the features of artificial mechanisms—that is, machines. I do not intend my use of the term "machine-like" to commit me to any view that sets me in opposition to Craver on this point. As I see it, artificial mechanisms will be the most machine-like systems, and biological mechanisms will be less so. However, and I return to the point made in the previous paragraph, the

question is whether some systems are so unlike machines in their organisation that the mechanistic approach to explanation loses its epistemic value and other forms of explanation should be used instead.

With these caveats out of the way, let me describe one framework for categorising systems as more or less machine-like. This framework comes from the work of Derek Skillings (2015), who constructs a three-dimensional space comprised of dimensions along which systems and processes can be more or less machine-like. These dimensions also reflect strands of criticism found within the philosophical literature. This space with Skillings' examples is shown in figure 4.1. The dimensions and those examples are explained below.

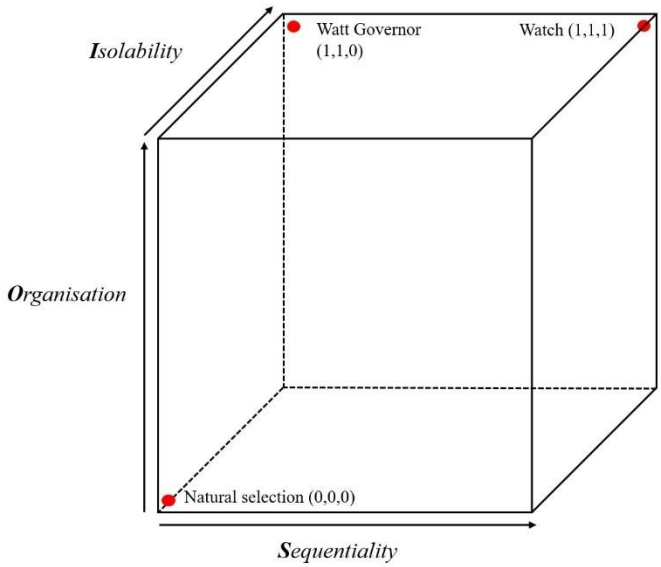

Figure 4.1 Skillings' three-dimensional framework for categorising processes that are more or less paradigmantically mechanistic. Adapted from Skillings (2015) and shown with three examples. The watch, a paradigmatic machine, scores high on all dimensions, while natural selection, a paradigmatic population phenomena, is a process scoring low on all dimensions. These are both Skillings' examples. I have added Watt's Centrifugul Governor as an example of another classic machine, but one that scores very low on sequentiality. This machine, and my reasons for including it, are discussed below.

The first dimension of figure 4.1's cube, is that of sequentiality. The motivation for this dimension derives from the characterisation of mechanistic descriptions found in Peter Machamer, Lindley Darden, and Carl Craver's influential paper on the subject, where they state that mechanistic descriptions typically begin by stating "start or set-up conditions," end with "finish or termination conditions," and contain a series of intermediate steps which are often "treated for convenience as a series of discrete steps of stages" (2000, pp. 11–12). Although Machamer and colleagues are careful to state that such descriptions are typically idealisations, William Bechtel and Adele Abrahamsen argue that natural processes so infrequently follow a steady and step-wise sequence that we should be sceptical of the value of such idealisations (Bechtel, 2011, 2012; Bechtel & Abrahamsen, 2005, 2013). While these idealisations might be helpful during the early stages of research, they argue, a more complete picture of a mechanism will require representing the more complex and often cyclical dynamics that biological systems and processes frequently exhibit. Among other things, differential equations, which are well-suited to describing dynamical systems, can be used to more accurately reflect the dynamics of activities in natural mechanisms. Returning to Skillings' cube, the sequentiality dimension tracks how well systems correspond to the step-wise and sequential ideal or a highly cyclical and dynamic alternative.

The second dimension is that of organisation. According to Skillings, this represents "the orderliness of the system in terms of the distinctiveness of its component parts, the discreteness of its component parts, and the importance of local relations between parts in producing the higher-level phenomena" (p. 1149). Since I have spoken, and will continue speak, about organisation in the terms introduced in section 4.2—that is, in terms of active, spatial, and temporal arrangement—I will rename this second dimension "orderliness" to avoid confusion while keeping faithful to Skillings' own description just reproduced. Inclusion of this dimension is motivated by Levy's and Bechtel's criticisms. They argue that the mechanistic account fails to distinguish between cases in which parts are functionally distinct and local interactions matter and ones in which parts are uniform and local interactions are unimportant (Levy, 2013, 2014; Levy & Bechtel, 2013). In making these criticisms,

Levy and Bechtel draw upon William Wimsatt's (1986, 2006, 2007) work on emergent and aggregate properties, which I will summarise now.

Wimsatt (2007, pp. 280–281) defines aggregate system-level properties as ones that meet four conditions: (1) the system property is invariant under rearrangement or exchange of parts; (2) quantitative properties scale with size increases and decreases; (3) the system property is invariant under the decomposition and reaggregation of parts; (4) and there are no cooperative or inhibitory interactions between the parts. Mass is a paradigmatic aggregate property. No matter how we rearrange the components comprising my body, decomposing and putting me back together, I will have the same mass. Likewise, increase my size and my mass will increase along with it—there will be no unexpected changes in mass scaling. Finally, the parts of which I am composed do not interact to produce my mass. Mass, as a system-level property, is merely the sum of my parts. Levy and Bechtel argue that, in many cases, components are aggregates rather than functionally integrated systems of functionally distinct parts. The dimension of orderliness in Skillings' framework, then, captures the range of cases from disorderly aggregates to highly ordered integrated systems.

The third dimension is that of isolability. This is based on the observation that processes can vary with respect to how easily we can draw a clear boundary around the process. Unlike Skillings' other dimensions, this one is not motivated by a specific criticism of the mechanistic approach found within the literature. Indeed, to the extent that low isolability makes a system or process less apt for mechanistic explanation, I suspect that these problems collapse into ones related to disorderliness. That is, if it is difficult to draw boundaries around a system or process, then it is unlikely that the system is highly functionally integrated. Nevertheless, it seems as though the extent to which a system is isolable is informative about the extent to which the system is a real, objective feature of the world. If a system is not isolable and has no boundaries, it's not obvious that there is a system or mechanism there at all.

Skillings also raises a further line of criticism found within the literature despite not allocating a corresponding dimension to it in his framework. This line of criticism regards the regularity of

component activities. Addressing the use of movie-like representations of ribosomes, Peter Moore (2012) argues that the mechanistic account is mistaken to suppose that entities perform their activities regularly. Molecules, for example, exist in a highly stochastic world and are at the mercy of forces such as "atomic bonds, friction, thermal forces, and internal vibrational decay" (Skillings, 2015, p. 1146). Consequently, the operations performed by molecular mechanisms are regular in idealised descriptions alone (see also Matthewson, 2017). Given this, another dimension could be added to Skillings' framework—regularity—to capture the diversity among processes and system with respect to the regularity or stochasticity of their activities.

Now, the concern is that the mechanistic approach is not apt for application when dealing with populations because populations score low on these four dimensions. As I see it, the dimension of orderliness is the most important for determining the applicability of the mechanistic approach. First, and as I said above, I suspect that the isolability of a system presents a problem for the mechanistic view primarily because of the constraints it places on orderliness rather than directly. However, I recognise that it is difficult to describe the component parts of a system if it's unclear where the system begins and ends. Second, I do not think that low sequentiality presents much of a challenge for a mechanistic approach. In part, this is motivated by classical machines such as the centrifugal governor, which score low on sequentiality but, as machines, are perfectly amenable to mechanistic analysis. The centrifugal governor is a piece of steam-age technology familiar to many philosophers (thanks in no small part to Van Gelder, 1995, 1998; Van Gelder & Port, 1995), which regulates the speed of a steam engine. The governor keeps the engine steady by connecting a pair of flyball arms to the engine and to a valve. The position of the arms is determined, via centrifugal force, by engine speed: higher engine speeds cause the arms to rotate faster and lift. Arm position, in turn, determines the degree to which the valve is open or closed, and the valve's position controls the speed of the engine via the amount of steam it allows through. Philosophers have used this machine as an exemplar of a dynamical system (Chemero, 2009; Van Gelder, 1998), which would score low on sequentiality because of the cyclical and continuous connection between steam input, engine speed, arm position, and valve position. However, it is my contention that a mechanistic analysis of the centrifugal

governor is unproblematic because the machine scores high on orderliness, which I take to be the most important dimension for determining the value of taking the mechanistic approach toward a system.

Why think that orderliness is the most important dimension for determining the value of taking the mechanistic approach toward a system, including a population? Let's return to the contrast Godfrey-Smith makes in the quotation included at the beginning of this section. There, Godfrey-Smith provides a clear statement of the differences between populations and "machine-like" systems, describing something called "population thinking," and suggests that this mode of thinking is importantly different from that of a machanistic approach. The features that Godfrey-Smith cites when making this distinction pertain to the orderliness of a system, as opposed to isolability, sequentiality, or even regularity (2009a, p. 148): "A highly structured network with heterogeneous and non-interchangeable parts is a different thing from a population." Likewise, Wimsatt takes features of orderliness to determine which systems are suited for mechanistic thinking (2007, p. 177): "Mechanistic explanations of phenomena commonly involve highly differentiated parts and behaviour that depend on their mode of organisation." Similarly, Levy and Bechtel are explicit about the effects of low orderliness on the value of mechanistic thinking (2013, p. 244):

> "It is the fact that the system is organized (and the type of organisation it has) that makes it amenable to mechanistic description and analysis. Systems that are internally disorganized—like a flux of diffusing molecules—where constituents do not make distinctive contributions, or internal integration is secondary, are ones in which decomposition tends (epistemically) to be less powerful."

The epistemic value of decomposing a system into its parts and charting their arrangment decreases as individual parts and their operations make less of causal contribution to the phenomena resulting from their interaction. In such cases, knowing about each of the parts, what they are doing, and how they are interacting, does not provide a great deal of counterfactual information. Rather, it is epistemically more powerful—that is, we can gain more counterfactual information—by treating the population as a whole. As Wimsatt argues in his work on aggregate phenomena, if it's possible to treat

systems as aggregates without disrupting the dependencies of interest to us, then we should do so, as it will simplify our representations and theories while focusing out attention on the key relationships responsible for producing a phenomenon.

Before looking at two example ABMs and presenting my positive view of how the mechanistic framework applies to populations and population thinking, I will consider an alternative view of the relationship between mechanisms and populations from the literature. I disagree with this view, but my reasons for disagreement highlight the unique representational features of ABMs.

4.4 A possible solution?

One line of response to the objection that the mechanistic approach to explanation has little value for populations is that population phenomena *can* be explained mechanistically, but only via another model which is far more machine-like than the population. I attribute this argument to John Matthewson and Brett Calcott, who present the argument in their (2011) paper, but Weisberg (2014) also appears to hold this view. In this section, I will present Matthewson and Calcott's argument along with the examples from their paper and Weisberg's. I will then present my counter argument to the view, which is that the "machine-like models" described by Matthewson and Calcott and Weisberg do not support mechanistic explanations as I described them in section 4.2. This is because these models and accompanying explanations do not describe the system's underlying component parts and their structural features. Rather, they describe causal relationships between the system's population-level properties. While these might be causal explanations, they are not mechanistic explanations.

Let's start with Matthewson and Calcott's argument. In their (2011) paper, the pair emphasise the distinction between models and targets, noting that each can have different properties. Given this, both can vary independently with respect to their degree of machine-likeness. Consequently, modellers can produce representations that are far more machine-like than their targets. Moreover, these representations can demonstrate dependencies that hold in the target and, by doing so, explain target phenomena resulting from these dependencies. Matthewson and Calcott then argue that these

explanations are mechanistic if two conditions hold: (1) the dependencies are present in the model in virtue of the internal structure of the model; and (2) these same internal relations can be found in the target. It is this final step with which I disagree. If they are right, then we might be able to produce mechanistic explanations of population phenomena, but it won't be with ABMs.

Let me make Matthewson and Calcott's line of argument concrete with an example. To scaffold their argument, Matthewson and Calcott use the Newlyn-Phillips Machine as an exemplar.[39] This model is also known as "MONIAC", a play on (1) "money" because the model represented the economy; (2) "maniac" because the model was very loud and very alien; and (3) "ENIAC," which was the first general-purpose digital computer. The Newlyn-Phillips Machine is an analogue computer built from water pipes, tanks, and valves, using water to represent quantities of money flowing through a national economy. The Machine is completely unlike any economic model that preceded or followed it.

While at the LSE, Phillips penned a paper which he asked Newlyn, his friend and mentor, to read (M. S. Morgan, 2012). At the time, Newlyn was only a year ahead in his studies at the LSE but was already lecturing at Leeds. In Phillips' paper, Newlyn found common economic diagrams redrawn as plumbing schematics. These new models were designed to demonstrate the dynamic quality of the economy lacking in the simple curves that most economists used to represent their ideas. Newlyn was especially captivated by the opportunity to explicitly represent the time-lags seen in real economic processes, having also played with the possibility of using pipes to demonstrate the structure of the economy in his own work. Newlyn and Phillips set to work together and, in 1949, the pair presented their working prototype to the LSE's economics department. Over the next few years, Phillips built a second version of the model, which was sold to various institutions such as Cambridge, Harvard, and

[39] While the first incarnation of the Machine is known as the Newlyn-Phillips Machine, the second incarnation is known as the Phillips-Newlyn Machine, reflecting the changing degree of involvement of the two collaborators. For the sake of brevity, and to refer collectively to both models, I will simply speak of the "Newlyn-Phillips Machine" or "the Machine."

Melbourne Universities (Morgan 2013, p. 178). The best, and perhaps only, place to see the model in the Southern Hemisphere today is in Phillips' home country of New Zealand, where one belongs to the reserve bank.[40]

The relevant features of the Machine for the present line of argument is that the target process is a population phenomenon, while the Machine, as the name might suggest, is paradigmatically machine-like: it *is* a machine. In terms of the dimensions described in section 4.3, the only one along which it scores low score is that of sequentiality. This is because, once the Machine is switched on and the water begins to flow around the system, there are no discrete steps and there are no clear stop conditions. Otherwise, the Machine is highly isolable—it is easy to draw a boundary around the system—and the system is highly ordered—each of the components is easily distinguishable as either a pipe or reservoir representing a flow, like taxation, or a store, like national savings. In addition to their distinguishability, the particular actions and interactions of these components matter, such that, if any of these components were unable to perform their designated functions, the Machine would fail to function. By contrast, an actual national economy is not so easy to isolate, nor are the components so easy to distinguish, and the actions and interactions of the components of an economy typically only make a difference in the aggregate with the exception of some institutions like reserve banks.

Explanations of economic phenomena provided by making interventions on the Machine, according to Matthewson and Calcott, will be mechanistic. This is because, although economies are not machine-like, the representation providing the explanation is machine-like. They remark that, importantly, the Machine does not provide mechanistic explanations because it is, in fact, a physical machine, although this makes its machine-like organisation obvious. The explanatory irrelevance of the physical properties of the Machine become yet more obvious when we consider, returning to the simple similarity view of scientific models presented in Chapter 1, that the Machine's pipes, tanks, and valves, are just a means of specifying an abstract structure—that is, a set of mathematical relationships

[40] The Reserve Bank of New Zealand also has a neat demonstration video accessible on YouTube: https://youtu.be/rAZavOcEnLg.

already known to economists at the time. Although it is undeniable that the visual aspect of the Machine, showing water moving through the system, filling and emptying from tanks, is part of what makes the Machine appealing, it is also worth recalling the discussion of visualisation from Chapters 1 and 2: visualisation is an important tool in computer simulation more generally. Visualisations can assist researchers when attempting to get a grip on a model with a complex structure, and things are no different here.

Matthewson and Calcott argue that the Machine supports explanations because we can make interventions on the model structure (via its specification, just like any other mathematical model) to produce outcomes that are sufficiently similar to outcomes produced by making the corresponding (mostly hypothetical) interventions on the model's target system. They take these explanations to be mechanistic because the Machine responds as it does in virtue of a set of relationships holding that are sufficiently similar to causal relationships that hold in the target, such as relationships between household income, savings, foreign debt, taxation, and so forth. I disagree that this is a sufficient condition for an explanation to be considered mechanistic. Before making my counter argument, however, let me report Weisberg's view of the relationship between mechanistic explanations and population phenomena. In doing so, I will show that Matthewson and Calcott are not alone in possessing their view of mechanistic explanation of population phenomena.

Weisberg (2014) describes an agent-based model of beech forests, where the model is used to investigate the formation of patches among otherwise dense forest. Weisberg argues that the model can support mechanistic explanations with the assistance of another model, which tracks the relationships between population-level properties. In terms of its structure, this second model looks much more like the relationships specified by the Machine than it does the ABM. Let's start with a brief description of the ABM before describing this further model.

The beech forest model is comprised, first, by a two-dimensional space with grid cells representing $14m^2$ of the forest. A vertical dimension 4 cells high then represents tree development and height, including a seedling layer representing the first 30cm, a juvenile tree layer spanning 30cm-

20m, a lower canopy layer spanning 20-30m and an upper canopy layer spanning 30-40m. Different numbers and kinds of trees can occupy these cells over time based on a set of rules. Relevant to the present discussion is that, from this ABM, Weisberg constructs a causal graph linking properties of the model, including "wind damage," which has a causal influence on other factors, like "neighbourhood interactions" and "diffuse and oblique light." It is this causal graph that Weisberg takes to support mechanistic explanations of forest patterns. And, according to Matthewson and Calcott, he would be right.

The causal graph supports counterfactuals that hold in the model due to its internal structure and which would hold in the target, supposing the model is accurate, in virtue of the same relationships existing between the target's properties. Weisberg states that mechanistic explanations require learning "about the counterfactual dependence of higher-level properties on lower-level mechanisms" (p. 789), though he doesn't state what he takes "mechanisms" to be. When discussing his causal graph, he states that the "causal graph of the [beech forest] model is clearly the representation of a mechanism. It shows the dependence of horizontal and vertical forest structure on neighbourhood interactions and mesolevel properties such as wind damage and incident light" (p. 792). While Weisberg is surely right about what dependencies the causal graph shows, an important step has been taken away from the characterisation of mechanistic explanation as one where system-level behaviours are explained by reference to underlying component parts and their active, spatial, and temporal organisation.

I am not convinced by the arguments and claims of Matthewson and Calcott and Wiesberg. Although they have provided convincing descriptions of one way in which we can produce models supporting causal explanations of population phenomena, these models do not support *mechanistic* explanations. While causal graphs or sets of differential equations solved through hydrological analogue computation are powerful representational devices for demonstrating counterfactual dependencies, they fail to possess the key characteristics of mechanistic explanation. To put it simply, an interventionist conception of cause and causal explanation, even combined with descriptions of lower-level properties, is not sufficient for mechanistic explanation. Indeed, the sorts of models

Matthewson and Calcott and Weisberg describe may very well support kinds of explanations-how, rather than mere what-but-not-how explanations. However, recall from section 4.2 that functional explanations are a kind of explanation-how that is distinct from mechanistic explanation.

Describing the parts or components that comprise a system along with their active, spatial, and temporal organisation is an essential feature of a mechanistic explanation, distinguishing it from causal explanation more broadly construed. In the case of Weisberg's graph, for example, while it represents the properties of a system and their influence on one another, these are not descriptions of the components underlying the system nor the activities, interactions, arrangement, or organisation of the underlying parts. In the case of Weisberg's causal graph, some of the nodes of the graphs do not represent properties but instead represent processes. "Neighbourhood interactions" and "vertical light competition" are processes that occur within the model. Although these processes are represented as distinct within the causal graph, they would all be mapped onto the same thicket of component parts because the same trees underly both neighbourhood interaction and vertical light competition. "Mortality rates" and "diffuse and oblique light," on the other hand, are not processes at all but higher-level properties. This model, then, fails to represent underlying parts and their organisation in the relevant sense. The same is true for Matthewson and Calcott's exemplar. Although the Machine represents the structure of the economy, it does so in terms of mathematical relationships between population-level properties rather than in terms of the organisation of physical components. This is insufficient as far as mechanistic descriptions go. As Craver states (2007, p. 162): "Organisation is not merely a matter of being describable in terms of a box-and-arrow diagram or program. Instead, it involves the active, spatial, and temporal organisation of different components."

A distinction drawn by Jaakko Kuorikoski in their (2009) paper is helpful here. They argue that, by "a mechanism for a phenomenon" we could have in mind (1) a *componential causal system* or (2) an *abstract form of interaction*. If we are talking about a componential causal system, then we are talking about decomposing a system into parts, which are often localised, and identifying the operations or activities played by these parts. If we are talking about an abstract form of interaction, then we need not be interested in parts, but can be interested in population-level properties such as

mortality rates or population-wide processes like light competition, which are sometimes difficult to map directly onto constituent parts and which "interact" with one another in so far as they satisfy Woodwardian conditions of causal influence—that is, interventions on *x* can produce changes in *y*. Such instances of abstract interaction are mechanisms only in the broader sense of a "causal driver," mentioned in section 4.2, rather than in the sense of the new mechanistic philosophy.

While I think Kuorikoski is right that we could mean either of these things when we say "a mechanism for," when it comes to mechanistic explanation, as opposed to, say, functional explanations or other kinds of causal explanations, we are talking about componential causal systems. Importantly, this is not to say that explanations referring to abstract forms of interaction—that is, causal dependencies which are not straightforwardly mapped on to component parts—are not good explanations or that mechanistic explanations are superior. Far from it. The point is simply that mechanistic explanation is a distinct kind of explanation, and that mechanistic explanations are distinguished on the basis of including information about component parts and the relationships between them rather than information about properties and the relationships between those.

This is not simply a matter of terminology either. As mentioned at the end of section 4.2, mechanistic explanations can be important in the context of intervention, since a researcher needs to know what part of a system they must manipulate to produce a desired result. One may very well wish to intervene upon a higher-level property, but such interventions can only be made via the entities from which these properties depend. Economists discussing ABMs today, for example, praise their capacity to represent the actions of individuals from the bottom-up (Turrell, 2016). This is a feature of mechanistic explanation and a feature completely missing from the Machine.

In this section, I considered an existing account of the relationship between population phenomena and mechanistic explanation. According to this view, mechanistic explanations can be provided via models that are more machine-like and more ordered than the populations they represent. The model will be explanatory, so the story goes, if the model explains in virtue of structural similarities between itself and the target. I argued that, although such models support causal

explanations, they are not mechanistic explanations because they fail to describe any interacting components or their active, spatial, and temporal organisation. These are the key features of mechanistic explanations.

In the next section, I will argue that ABMs can represent populations and their components and support explanations where the arrangement of parts, or the orderliness, provides explanatory power.

4.5 Applying mechanisms to ABM

Many proponents of ABM describe ABMs as mechanistic models, including the authors of the first example to be shown below (Day, Zollner, Gilbert, & McCann, 2020, p. 2193): "Individual-based models (IBMs) provide a mechanistic perspective on the patterns observed in ecological systems, including the estimation of landscape connectivity." Is this just loose talk or is there a deeper connection between ABM and the mechanistic approach? In this section, I argue that ABMs decompose populations into their constituent parts to investigate how the actions and interactions of those parts contribute to the phenomenon of interest. ABMs also permit the explicit representation of active, spatial, and temporal organisation or parts and their activities, including at multiple levels or scales of organisation. Furthermore, the system-level patterns investigated by modellers sometimes rely on the specifics of this organisation, meaning that removing the description of organisation would produce models of lesser value relative to the modeller's research questions. These features of ABM demonstrate the applicability of the mechanistic approach to ABMs and population phenomena. When modellers refer to their ABMs as being mechanistic, there is more to this than loose talk.

To make my argument, I will present two example ABMs in sections 4.5.1 and 4.5.2, which I will analyse in section 4.5.3. I will begin by making three quick caveats regarding these examples.

First, you may recall from Chapter 1 that I have already introduced an ABM: Jones' virtual slime mould. Why introduce different models here? For now, I wish to concern myself with

straightforward cases of virtual populations representing population phenomena, setting the virtual slime aside as it is a representation of a single organism rather than a population. I will briefly return to Jones' model in section 4.6, where I discuss the explanatory power of ABMs. There, I will consider the benefit of representing physical systems, materials, or other non-population systems as populations.

Second, and regarding the quality of the models I have chosen, let me emphasise that these models have not been chosen because I think they are good models relative to their purposes. Rather, I have chosen them because they are good models relative to *my* purpose—that is, demonstrating the prima facie applicability of the mechanistic framework to ABMs and population phenomena. For example, spatial organisation and even the physical structure of some entities matters within these models relative to the production of the target phenomena, at least to some extent. In other ABMs, space does not matter at all, and many ABMs do not explicitly represent space or the spatial arrangement of parts. In choosing models that exemplify my claim about the applicability of the mechanistic approach to ABM, I am not problematically stacking the deck in my favour. This is because, by the end of the chapter, I will have made the case that there is a gradient of cases. Here, I just wish to present cases that sit around one end of the gradient.

Finally, given that a detailed criticism of the models is unnecessary for my argument, I will attempt to avoid including extraneous details in my presentation of the examples, describing only the core features of the models in order to explore the applicability of the mechanistic framework to populations.

4.5.1 Example 1: A model of dispersal

The first example is an ecological ABM used to explore the effects of mortality, asymmetric dispersal, and land-use change on the functional connectivity of populations of the American marten (*Martes americana*), a member of the weasel family, in Wisconsin and Michigan (Day et al., 2020). Casey Day and colleagues focus on American martens because they were once eliminated from these areas but

have been restored due to conservation efforts. Ongoing conservation efforts, then, would be assisted by such a model. The theoretical motivation behind the model is that connecting otherwise isolated populations together or to larger population groups via dispersal, known as functional connectivity, increases the chances those populations will persist. However, most models investigating connectivity do so by representing the systems from the top-down rather than from the bottom-up, leading them to omit certain features that Day et al. argue make a difference to functional connectivity in the real world. Three typically missing features that Day et al. include in their model of dispersal include (1) mortality risks from starvation, predation, or human interference, (2) asymmetric dispersal, and (3) habitat fragmentation caused by land-use change.

Two kinds of entities included in the model are the virtual environment and the virtual dispersers. The virtual environment represents a 2502 km^2 portion of North West Wisconsin, extending somewhat into Michigan. This area contains a mixture of forested and non-forested land, with the forest serving as a corridor between two reintroduced populations to the east and west respectively. Within the model, the environment is comprised of four spatially explicit maps that contribute to disperser behaviour, representing movement possibilities, food availability, predation risk, and habitat availability. Dispersers seek to establish a home-range location within the environment that is unoccupied by a member of the same sex, with dispersal broken into an exploration phase and exploitation phase. During exploration, dispersing individuals interact with the maps just described and keep a log of potential habitat locations. During this phase, dispersers may also die if they starve, fall victim to predation, or fail to find an acceptable home-range. After a period of time set by the modeller, the individuals switch to exploitation, changing their movement from a semi-random walk to heading directly for their chosen site, with algorithms in place to control their behaviour if no suitable location has been discovered during dispersal (keep searching) or the chosen site is occupied by another before reaching it during exploitation (head to the next best site in the log).

Moving on, now, to results of the model study and setting aside any further detail of model processes, how did the three factors of mortality, asymmetric dispersal, and land-use change impact functional connectivity? Mortality was found to have the biggest negative impact on the ability for

successive generations of dispersers to cross the corridor and establish a home-range in the territory of the opposite side's population. For Day et al., this demonstrates that other researchers may have misjudged the effects of habitat fragmentation on functional connectivity. This is because the omitted travel costs associated with moving through non-forested areas lead to an overly optimistic view of the impact of fragmentation on functional connectivity. Perhaps counter-intuitively, Day et al. demonstrate that habitat loss simpliciter may be less important for functional connectivity, and hence population persistence, than habitat fragmentation. With their model, Day et al., experimented with two possible placements of a proposed mine. They found that placing the mine in an already fragmented area would have a larger negative impact on functional connectivity, despite removing less habitat, than if the mine were placed in an area of greater contiguous forest. Once again, this is because of the mortality risks associated with passing through fragmented habitat.

4.5.2 Example 2: A model of polity recycling

The second ABM is a representation the Lake Titicaca basin, now straddling the borders of Peru and Bolivia, and the development of political entities within this area from about 2500 BC to about 900 AD (Griffin & Stanish, 2007). The purpose of the model is to recreate four patterns observed in the archaeological record: (1) cycles of fission and fusion of political entities; (2) population concentration at primary and regional population centres; (3) primary population centres located at Lake Titicaca's north and south; and (4) population growth rates of 0.1% reaching a maximum of about 500,000. By recreating these patterns via the representation of much smaller chunks of the population and their political activities, the model is intended to help explain those patterns in terms of the actions and interactions of those lower-level entities.

These patterns from the archaeological record are used to determine the realism of the model but are further broken down and split into twelve criteria. These include, for example: that the final population is close to its expected value (related to the fourth pattern above); that the north polity is

larger than the southern polity before the collapse of Pucara, a regional polity, around 200-300 AD; that the southern polity is twice the size of the north polity after the collapse of Pucara, and so on.

The model includes five entities: Patches, Settlements, People, Chiefs, and Polities. A 200 by 200 grid of Patches, which represent an area of 1.5km^2, comprise the virtual Lake Titicaca basin. Using detailed maps, Patches are sorted into geographic zones based on hydrology and elevation, such as "lake edge," "swampy," or "hillside (3900m and 4000m)" zones. Based on inferences from the archaeological record made within the literature, Patches are assigned possible agricultural uses. These possible agricultural uses are based on geographic zones, access to water, and whether political complexity enables high intensity or low intensity agriculture. Settlements occupy Patches and undertake a range of actions, such as expanding to new Patches, creating new Settlements, abandoning Patches, planting and harvesting crops, and following or resisting the Chief in charge of the polity to which they belong. Settlements are inhabited by People. People do not represent individuals but are used to track changes in population numbers as well as location. The ability for parts of a population to move between settlements motivates the representation of People and Settlements as separate entities within the model. Chiefs are centred on one Settlement each and control a Polity, which contains all the Settlements led by the same Chief. Chiefs take a share of the harvest from Settlements within their Polity and can move their centre if another location would increase their strength, which is determined by the population of the Settlements in their Polity, modified by the distance of a Settlement from the Chief's centre. Chiefs can also attempt to control Settlements of other Chiefs and assert control over any Settlements within their own Polity attempting to resist their leadership, the success of which are both determined partly by the Chief's strength. As these last two actions of the Chief suggest, neighbouring Polities can merge into larger Polities or Polities can dissolve into independent Polities. For all entities that make decisions, their decisions are based only on information regarding neighbours; they do not have any global information. So, Chiefs have information about

Chiefs in neighbouring Polities, People have information about neighbouring Settlements, and Settlements have information about neighbouring Patches.[41]

The key processes, repeated at each time step, are a series of operations, mainly based on decisions by Chiefs and Settlements, which end with a possible restructuring of Polities. This starts with the Chief deciding whether to shift his centre or stay put, Settlements planting crops or establishing herds and harvesting a yield, and Chiefs taking their share of a harvest. Settlements then decide whether to perform one or none of the following actions: "colonizing new Patches, spawning a new Settlement, abandoning the Settlement, or seceding from a complex polity" (p. 34). Afterward, Chiefs decide whether or not to initiate a conflict either internally, which involves supressing rebellion from a seceding Settlement, or externally, which involves attempting to gain control over a neighbouring Polity. The results of any conflicts are then calculated, and the population of People is adjusted according to births, migration, and conflicts. Finally, Polities are adjusted based on the outcomes of conflicts.

I have omitted many details of this model for the sake of brevity, but we should now have enough to move to the results of the model. Although not all 12 patterns mentioned earlier were reproduced by each run—indeed, the runs that did were in the minority—about a third of the model runs resulted in matching 11 out of the 12 patterns. Moreover, the authors of the study calculated that, even though runs meeting all 12 criteria were in the minority, the model was still 60 times more likely to produce this result than chance, which they take as an indicator that the model has explanatory power and relevant similarity.

Here is one explanation that the authors derive from their model study. This explanation regards the emergence of political and population centres at the Northern and Southern points of the lake, which they attribute to the geography of the lake. Put simply, the narrow northern and southern

[41] Unfortunately, Griffin and Stanish (2007) fail to state which kind of neighbourhood is used in their model. However, careful ocular inspection of the "Animation of the simulation run shown in Figure 4 of text" suggested a Von Neumann neighbourhood.

points of the lake allow for a greater concentration of nearby populations compared to the west and eastern axes of the lake. There, less space is available for settlement and agricultural use, contributing to lower populations and correspondingly lower strength political entities. Additionally, a Chief's access to trade passes, which are located at a series of Patches to the East and West of the lake and which are intended to represent the connection of the Basin region to communities further afield, increase his strength. Chiefs located along the Western or Eastern banks of the lake can, at most, hope for access to Eastern or Western trade passes, but not both. The authors offer other explanations for observed patterns with their model, but this one will suffice for my purposes.

4.5.3 Scoring the examples

With the two examples on the table, we can now consider whether these systems benefit from being approached mechanistically. To start, let's score these cases along the dimensions of Skillings' framework. There will be no surprises here, as the models and their targets both tend to score lower than higher on the dimensions of machine-likeness.

First, sequentiality. In the case of American Martens dispersing across habitat corridors, the process is not highly sequential, though there are some key steps. Perhaps most obviously, there are seasons which impact when the critters move and when they need to find shelter. However, dispersal is mostly a continual process as individuals come from their eastern or western populations and slowly move around the area and (sometimes) across the corridor. This is reflected within Day et al.'s model, with the model process broken into two phases of exploration and exploitation which might be thought of as mirroring the seasons (warmer for exploration and colder for exploitation). To reflect the constant dispersal, whenever an agent dies, they are replaced by a new one spawned at one of eastern or western population territories. In the case of polity recycling around Lake Titicaca, things appear slightly more sequential. In the model, of course, there is greater sequentiality than in the target, as everything runs according to a set procedure which is absent in the target. For example, there is a particular order in which everything occurs at every time step—first Settlements harvest, then Chiefs

take their share, and so on—which is unique to the artificial conditions of the model. While it's true that, in the real world, harvests are collected at relatively set times as dictated by the seasons and particular life cycles of the crops, and it's true that the Chief cannot take their share before it is harvested, there is obviously no strict order of operations. Some of the patterns that comprise the explanatory target do emphasise sequentiality of larger-scale trends, such as the early northern centre and later more stable southern centre, and Griffin and Stanish's explanation of this does revolve around the order of events, such as the slow population growth in the south.

Second, orderliness. Starting with the Marten population, if we consider how functionally distinct the parts are, we find that they are relatively functionally uniform. While this is certainly true of the population members, it is slightly less true of the habitat, which is less uniform, including areas where mortality risks are higher, where there are more predators, where interactions with humans are more common, and where food is scarcer. As we can see from Day et al.'s explanation of the importance of mortality risk, these functional differences are crucial in the resulting connectivity between the populations on either side of the corridor. Furthermore, local interactions matter within the target and the model, and rearranging the parts could lead to radically different results given the impact of habitat fragmentation on mortality and, consequently, functional connectivity. In the case of the polity recycling phenomena and model, there appears to be a higher degree of orderliness because we are dealing with social hierarchies. Chiefs are examples of single individuals with very different active powers than, say, a common person. So much so that individual people are not even represented in Griffin and Stanish's model, but individual chiefs are. Still, within the classes of entities, there is a large amount of functional uniformity. Nevertheless, it's clear that the social world represented with Griffin and Stanish's model lacks many of the features Wimsatt attributes to aggregates, such as the ability for system-level properties to remain unchanged under rearrangement of parts. In the case of the polity recycling model, exchanging and rearranging parts can have a huge impact on the evolution of the distribution and stability of polities and the patterns that model is able to reproduce. While far from a highly integrated system, the polity recycling case, both phenomenon and model, is more orderly than Marten dispersal.

Third, isolability. In both cases, there is an obvious sense in which the model is far more isolable than the real populations. For all models, the model systems will be highly isolated because they are constructed artefacts with all features specified and no more. However, this does suggest that targets are at least somewhat isolable or the extent to which they could be modelled as isolated systems would be highly constrained. In the Marten population, there is some isolability in so far as the range of the Marten population does not extend far beyond the forests and the eastern and western home ranges. However, there are some unclear boundaries with respect to the corridor because exactly where we draw the line separating it from the territories is somewhat arbitrary. Things are similar in the case of the Lake Titicaca population. Although the real population was connected to and interacted with groups situated outside the Lake region, the relatively high density and interconnectedness of the Lake populations make them relatively easy to isolate and represent as such in a model (although note that model also attempts to represent the causal connections to populations beyond the Lake).

Finally, regularity. In both cases, the real systems and the models are highly stochastic. Of course, there is more regularity in the model as, once again, it is an artificial thing performing according to a designed set of rules. However, those specific rules include stochasticity in an attempt to capture the irregularity of the target. For example, Griffin and Stanish explicitly note that there is a huge number of different possible trajectories that their model can take, only some of which resemble anything like what is found in the archaeological record.

As I've argued in section 4.3, low machine-likeness scores do not imply that a system is not amenable to mechanistic explanation. Unpacking what aspects of orderliness are present in the model and the explanation provides more insight. Consider the inclusion of parts, which were missing in the cases described in section 4.4. ABMs explicitly represent the entities comprising a population, producing population phenomena from the bottom up (Bankes, 2002; Turrell, 2016). As mentioned, this representational capacity gives them an edge, in the eyes of many proponents of ABM, over mathematical and statistical models that cannot represent these entities or their local interactions (An, Grimm, & Turner II, 2020). Irrespective of whether explicitly representing these parts gives ABMs advantages over mathematical or statistical models that omit them (presumably there is a mix of

advantages and disadvantages), this focus on, and explicit representation of, underlying parts definitely coheres with the mechanistic philosophy's emphasis on parts and their actions. When decomposing populations into their constituent parts, these parts can be single individuals as in the case of the first example, or collections of individuals as in the second model, where the sheer complexity and size of the system being represented prohibits explicit representation of each member of the population.[42]

Mechanistic descriptions include the explicit representation of not only component parts but also their organisation. Here too, ABMs appear to fit with the mechanistic philosophy. In terms of active organisation, relationships between entities are not primarily represented as dependencies holding between entire populations. Instead, they are encoded at the level of entities which are programmed to act and interact with other entities according to specific rules: Settlements are programmed to expand or rebel, Chiefs are programmed to conquer or quash. ABMs are also appreciated by modellers for their ability to represent actions that are less easily depicted through other representational forms. For example, Day et al. argue that the methodology of ABM allows them to represent an important action performed by dispersers, which is that they die under certain circumstances relating to the distribution of resources and threats, and that this contribution can be used for better conservation and ecological management.

In terms of spatial organisation, the explicit representation of space within the two examples demonstrates that the spatial arrangement of component parts can be represented with ABMs.

[42] It would certainly be possible to nest, within the Settlement entity, a model of Settlement decision-making which did explicitly represent the individual occupants; however, given the purposes of the model, it's not obvious that representing the population at the level of individuals would improve the model. As with all mechanistic descriptions, there is a point at which they bottom-out (Machamer et al., 2000). This is where further levels of decomposition would not assist with representing or understanding the system and phenomenon under investigation. In the first model, it bottoms out at the level of the individual animals and, in the second model, it bottoms out at the level of population clusters.

Moreover, these models can be used to investigate phenomena where spatial arrangement plays a part in the emergence of the focal phenomena. In the first example, the spatial arrangement of habitat, whether it is contiguous or fragmented, impacts the overall functional connectivity between populations. Interventions can also made with the ABM that affect its spatial arrangement, such as the repositioning of the proposed mine site, which result in changes to this system-level property. In the second example, the spatial features of Lake Titicaca and the constraints it places on Polities helps to explain why the Northern and Southern points of the lake become political and population centres: Settlements and Polities along the east and west banks of the lake have less room to occupy and poorer access to trade passes to the east and west.

Mechanists also emphasise that complex systems are frequently hierarchically organised, with parts decomposing into further collections of interacting components (Bechtel, 2016, 2017; Machamer et al., 2000). Consequently, providing a complete description of a mechanism often involves representing systems at these different levels of organisation. According to proponents of the ABM methodology, these models are likewise apt for representing systems at multiple levels of organisation (Smaldino, Calanchini, & Pickett, 2015). In the second example, for instance, multiple levels of organisation are included, from groups of People which inhabit Settlements, to Chiefs that control Polities, which are collections of Settlements.

Finally, in terms of temporal organisation, ABMs necessarily represent the evolution of systems through time. As before, ABMs are not only capable of representing the temporal arrangement of activities and interactions, but these temporal factors can have important and interesting consequences for the emergent patterns of interest to the modeller. For example, in the second example, Griffin and Stanish explain the dynamics of an early northern political centre followed by a later, more stable, southern political centre in terms of temporal arrangement and the factors determining the probability of fission. Within the model, the probability of Polity fission is a function of the probability of Settlement rebellion and the strength of the Chief. Probability of a rebellion within a Polity increases linearly with the number of the Settlements included in the Polity because the probability of Settlement rebellion is constant. But the chances that a Chief will be able to

quash a rebellion increase as the population within the Polity increases. According to Griffin and Stanish, the political instability of the northern region is caused by rapid Settlement expansion early, caused by favourable hydrological conditions in this region, which leads to fission. This is then followed by continued cycles of conflict and consequent population leveling, which prohibits a single Chief from establishing control of the whole area. In the south, the population and number of Settlements grows more slowly, with large consolidation being delayed until the point where, when it does occur, population structure is such that fission is unlikely because there a fewer Settlements with large populations rather than more Settlements with smaller populations.

4.5.4 Summary

The preceding analysis demonstrates the following. First, populations, including the virtual populations of ABMs, score relatively low on the dimensions of machine-likeness. But this is no surprise; they are populations after all. The question is whether this makes taking a mechanistic approach to explanation lack value. As I've demonstrated, this is not so. This is because factors regarding parts and the specifics of their active, spatial, and temporal organisation makes a difference to the explanatory target. Even though the systems are not highly ordered, these aspects of orderliness make enough of a difference to the behaviour of the system that they are worth describing to answers at least some questions we might have about the target. The Godfrey-Smith quotation presented earlier suggested that population thinking was antithetical to a mechanistic approach, but I've shown that there are cases where populations should be treated mechanistically—that is, information about their parts and the organisation thereof should be described explicitly rather than abstracted away. In the next section, I want to continue the exploration of the connection between population thinking and explanation by showing that there are multiple explanatory strategies we can take when approaching population phenomena and that ABMs are an approach to population phenomena bearing the features of mechanistic explanation.

4.6 Population thinking (at least) three ways

So far, I have described a puzzle: many modellers think of ABM as supporting mechanistic explanation, and some philosophers think the same. But, just as many philosophers, if not more, think that little is to be gained from approaching populations like more machine-like systems. The response I have given so far is that the machine-likeness of a system does not decide whether the mechanistic approach has any explanatory value. The machine-likeness of a system might tell us *something* about how valuable information about organisation will be, but it certainly does not imply that such information has no value at all. A system might not be very machine-like, but its parts and the specifics of their organisation might still be important enough that providing this information improves an explanation in some cases.

My aim in this section is to resolve any remaining puzzle about the relationship between population thinking and explanation, mechanistic or otherwise. To do this, I will demonstrate that population thinking has no set explanatory framework by showing that there are at least three ways that we can explain population phenomena. These different explanatory approaches reflect the differing explanatory importance of the explicit representation of parts and their active, spatial, and temporal arrangement. The last of these three is the one supported by ABM, and which puts the greatest explanatory weight on parts and organisation, though still lacks some key features of mechanistic explanation such as the physical structure of component parts.

4.6.1 Explanation 1: Population-level dependencies

The first way that we might explain population phenomena was described by Matthewson and Calcott and Weisberg in the previous section. That is, we can isolate dependencies between properties of the population which impact the phenomena of interest. Much modelling of population phenomena looks like this, especially during the period before the widespread availability of computational power. Consider, for example, predator-prey models from ecology, such as the Lotka-Volterra model.

This model can be described with the following equations:

$$\frac{dV}{dt} = rV - (aV)P \tag{4.1}$$

$$\frac{dP}{dt} = b(aV)P - eP \tag{4.2}$$

Here, we calculate the change in prey abundance by multiplying prey birth rate *r* and prey abundance *V* and subtracting the captured prey. We calculate the captured prey by multiplying a capture rate *a*, the number of prey *V*, and the number of predators *P*. It is, of course, for the sake of simplicity that we assume that the prey population is limited only by predator consumption rather than additional factors like age and disease. We then calculate the change in predator abundance by multiplying the number of captured prey (*aV*)*P* by a conversion rate *b* and subtracting the dead predators, which we calculate by multiplying predator abundance *P* by a deathrate *e*.

The Lotka-Volterra model explains a system-level property—the change in the proportion of predators to prey after the conditions of a general biocide—by referring to dependencies between populations and population-level properties.[43] Although there is some representation of component activities in the model—prey reproduce, predators catch prey—these activities are represented as population-level properties rather than component-level properties. There is also no explicit representation of the parts of the system below the level of entire populations, and there is no representation at all of the spatial or temporal organisation of the components. This kind of explanation of population phenomena is not at all mechanistic and involves representing systems as mere aggregates.

[43] That is, if the model explains at all. On the view offered throughout this thesis, a model is explanatory not only if it represents dependencies, but if it represents *actual* dependencies. Here, I offer the Lotka-Volterra model merely as an example of what an explanation of population-level properties while remaining at the level of population-level properties.

4.6.2 Explanation 2: Population-level algorithm

This next kind of explanation is based on certain characterisations of natural selection. For example, Godfrey-Smith (2009a, p. 19) refers to descriptions of natural selection, such as those offered by Richard Lewontin, as "recipes for change" (Lewontin, 1985, p. 76; c.f. Godfrey-Smith, 2009a, p. 18):

> A sufficient mechanism for evolution by natural selection is contained in the three propositions:
>
> 1. There is variation in morphological, physiological, and behavioural traits among members of a species (the principle of variation).
> 2. The variation is in part heritable, so that individuals resemble their relations more than they resemble unrelated individuals and, in particular, offspring resemble their parents (the principle of heredity).
> 3. Different variants leave different numbers of offspring either in immediate or remote generations (the principle of differential fitness).

An alternative presentation of this recipe for change can be found in Robert Skinner and Roberta Millstein's (2005) paper, where they break the process into seven stages and the causal connections between each stage. For the sake of brevity, we need not go into the details of their particular formulation of natural selection, but it follows a similar, though more detailed, pattern. Skinner and Millstein's work, however, is especially relevant to the arguments of this chapter as their paper directly engages with the debate over the applicability of the mechanistic approach to population phenomena. Specifically, they argue that natural selection *is* a mechanism, though one that is not captured by the new mechanistic philosophy, which they see as a problem for the mechanistic philosophy.

While Skinner and Millstein argue that the inapplicability of the mechanistic framework to natural selection is a problem for the framework, I disagree. This is because, as noted in section 4.2, not every instance of "mechanism" talk in science is an instance where the mechanistic approach to explanation applies. Rather, "mechanism" is often meant as the more general "causal driver." Skipper

and Millstein themselves recognise that natural selection is referred to in many ways in addition to "mechanism" (p. 328): "Natural selection is a 'cause,' a 'force,' a 'process,' a 'mechanism,' a 'factor.'" Of all these ways of talking about natural selection, I do not see a good reason for thinking that "mechanism" is the one that should be taken philosophically seriously. Indeed, Skipper and Millstein provide the argument in their paper for why philosophical accounts of mechanisms do not apply well to natural selection. Their worries include, for example: how to individuate parts if the parts include the environment (is the environment one part or many, and what's counted in the environment?); the fact that natural populations are not organised but dispersed and changing; and the fact that many of the activities performed by interacting and reproducing populations do not have set temporal properties like duration, but temporal properties that vary and can change year to year.

Rather than developing a yet more general philosophical account of mechanisms as Skinner and Millstein suggest, I believe we can dissolve the problem by recognising that natural selection is a mechanism only in the broader "causal driver" sense of the term and need not be captured by the narrower sense of the new mechanistic philosophy. When it does come to considering the framework of explanation that applies to natural selection, I suggest that we look to alternatives to the mechanistic framework. Following Godfrey-Smith's general description of these as recipes for change, I propose thinking of these as "algorithmic" explanations, which proceed by describing a recipe or algorithm—a series of steps—that, when followed, produces the phenomena to be explained.

Now, I have no intention of producing a general philosophical theory of algorithmic explanation here. That would be a project all of its own, which must be left for future work. One possible way of making such an account, which I will mention here, is to look to the existing frameworks of functional analysis, like that described by Cummins (1975) and contrasted with the mechanistic approach in section 4.2. Functional analyses can be directed toward systems, such as the imitation system (c.f. Heyes, 2018b), or they might be directed toward a process. What I am calling algorithmic explanations, then, might just be familiar functional explanations, but ones that focus on processes rather than systems. Once again, I do not intend to produce a general theory of algorithmic explanation. Rather, my purpose is to demonstrate another explanatory approach that may be taken in

the context of population phenomena, different to both the population-level dependency approach described in section 4.6.1 and the approach taken with ABMs.

To end this sub-section, I will note some of the features of algorithmic explanation in the case of natural selection with the purpose of comparing and contrasting it to mechanistic explanations and the dependency-based explanations of section 4.6.1. Instead of decomposing a system into physical parts or isolating relationships between population-level properties, algorithmic explanations proceed by breaking processes down into key steps. These steps are then followed—though they need not be consciously followed, of course—to produce some phenomena. Just as in population-level dependency explanations, physical components are included in algorithmic explanations because there must be some physical thing that is following the rules of the algorithm. For example, there are populations of organisms that vary with respect to some morphological, physiological, or behavioural traits. However, the abstract process and the stages of the process are the explanatory focal points rather than these parts and their activities.

As in population-level dependency explanations, the spatial arrangement of the physical components is not explicitly represented and frequently does not matter. Now, of course spatial factors do matter when applying the algorithm to actual populations because spatial factors can affect which individuals are included in the reproductive population and those that are excluded. This is true in the case of population-level dependency explanations too. However, explicit description of space and spatial factors is omitted from both Lewontin's and Skinner and Millstein's recipes for change. The omission of spatial information is expected since algorithmic explanation, like functional analyses more broadly, takes place at a higher level of abstraction than a complete mechanistic explanation. As Skinner and Millstein argue, temporal organisation at the level of individuals might not matter in an explanation of natural selection because there is so much variation in the duration and order in which individuals execute their activities. However, at the level of the algorithm's stages, temporal organisation is crucial. Though the duration of steps is not stated in either Lewontin's or Skinner and Millstein's recipes for change, the order of the stages or steps in the recipe is explanatorily critical. The importance of temporal order is something that sets algorithmic explanations apart from

population-level dependency explanations. That, and that population-level dependencies appear to describe systems rather than procedures.

Now, let's turn to the sorts of explanations which can be provided by ABMs.

4.6.3 Explanation 3: Individual-level explanation?

In this subsection, I argue that ABMs, at least spatially implemented ones, are mechanistic descriptions of population phenomena. As mechanistic descriptions, ABMs explicitly decompose a system into its constituent parts. Population members are always among the parts—indeed, sometimes they are the only parts included—and the environment can also be among the physical things explicitly represented. Also, in accordance with the mechanistic framework, ABMs include the explicit representation of part activities, which are often the rules driving the behaviour of population members and the behaviours those rules control. These can be rules like *if n% or fewer of your neighbours are not like types, then move to a random unoccupied space* as in the Schelling model of segregation (Weisberg 2013), *imitate the best* as in a possible model of social learning, the more complicated rules determining the actions of virtual martens in Day et al.'s model, or the even more complicated rules determining the actions of the subpopulations in Griffin and Stanish's model.

In accordance with the mechanistic framework, ABM enables the explicit representation of the spatial arrangement of parts. As I have mentioned, there are many ABMs that are completely non-spatial, so explicit representation of spatial arrangement is far from necessary. In such cases, for example, virtual agents might simply be paired at random, drawn from a mathematical hat, rather than interacting in a virtual space. Alternatively, some non-spatial structure might determine the bias in their interactions, such as in the case of a network of researchers. Nevertheless, in cases where spatial structure is explicitly represented, there are at least two good reasons to do so. The first reason to explicitly represent spatial arrangement is that the target phenomenon is a spatial pattern. This is the case for models of segregation or flocking: segregation and flocking are spatial patterns that the model is intended to reproduce. The second reason to explicitly represent spatial arrangement is that the

interaction is taking place in a heterogeneous space and the distribution of relevant features in that environment plays a large role in producing the target phenomenon. ABMs that explicitly represent space, especially those that attempt do so realistically, such as the models discussed in section 4.5, are similar to mechanistic descriptions in this respect.

Finally, the temporal arrangement of activities is another crucial part of the mechanistic framework, which can be represented with ABM. ABMs are (with one or two notable exceptions) investigated through the use of computer simulation. As Hartmann (1996; see also Humphreys, 2004) emphasised, simulation is dynamic, representing the change in a system over time with a physically implemented computational process that itself changes over time. Running simulations, like the evolution of real systems, takes time. Parker (2009a) reflects this dynamic aspect of simulation in her own definition of computer simulation as a time-ordered series of states that represents another time-ordered series of states.[44] Although, in Chapter 1, I resisted the view that temporal dynamics are an essential part of computer simulation and its epistemology, it is undeniable that computer simulation, ABM included, is perfectly well suited to representing temporal organisation. This capacity has also been demonstrated by section 4.5's case studies. The model of polity recycling in the Lake Titicaca region, for example, was explicit about temporal factors like the duration of polities, the rate of rebellion, and the order of power concentration from the north to the south. ABM, therefore, can include temporal arrangement in accordance with the mechanistic framework.

The representational capacities and uses of spatially implemented ABMs are sufficiently unique that I will distinguish between two kinds of ABM. In each case, a different aspect of the ABM carries the explanatory weight. In the first case, where ABMs are not spatially implemented, the factors carrying the explanatory weight are the rulesets that guide the agents' activities. In these cases, changing the behavioural rules can change the resulting population phenomenon. This ability to make interventions on a global pattern or property by making an intervention on the agent rule provides the rule and the activities it controls their explanatory power. In the second case, ABMs still include the

[44] This is only for a single simulation run. Many runs are typically used in a simulation study.

explicit representation of agents and their rules, but they also explicitly represent the spatial arrangement or sets of possible arrangements of these agents within a heterogeneous environment. In these cases, agent rules do not carry the explanatory weight or, at the very least, they do not carry all the explanatory weight, sharing it with the features of the heterogeneous environments with and within which they interact. In such cases, interventions on that heterogeneous environment can produce a counterpart change in the target phenomena. For example, placement of the mine in the model of American marten dispersal will make a difference to the functional connectivity of the two sub-populations: if it breaks up an already fragmented dispersal channel, functional connectivity might collapse, if dispersers can funnel around the mine site, then functional connectivity does not collapse.

Having distinguished between spatially implemented and non-spatially implemented ABM, I will now compare different models according to which aspects of the mechanistic framework they include. That is, whether they include the explicit representation of underlying entities, the activities of those entities, the spatial arrangement, and their temporal arrangement. A comparison chart is shown in table 4.1 and helps to explain why modellers and philosophers talk about ABMs in mechanistic terms. Of the representational approaches taken toward populations and reviewed in this chapter, only ABMs include the explicit representation of entities and interactions, organised so as to produce the target phenomenon. The population-level algorithm or recipe for change involves decomposition but not of entities, while the population-dependency model and explanation omit decomposition entirely.

	Mechanistic	**ABM with space**	**ABM no space**	**Population-level algorithm**	**Population-level dependencey**
Entities	Yes (heterogeneous)	Yes (uniform)	Yes (uniform)	Procedures (variables)	Variables

Activities	Yes	Yes	Yes	Yes	No (properties)
Spatial Organisation	Yes	Yes	No	No	No
Temporal Organisation	Yes	Yes	Yes	Yes	No

Table 4.1 Mechanistic descriptions include the explicit representation of entities, their activities, and their spatial and temporal organisation. ABMs can represent these same elements. However, ABMs can also omit the explicit representation of space and many do. These three can be contrasted with the models accompanying population-level algorithm and population-level dependency explanations, which do not decompose the system or, if they do involve decomposition, do not involve decomposition of the right thing.

The comparison in table 4.1 raises the question of what changes when the move is made from representing uniform sets of entities, as in ABMs and populations, to heterogeneous sets of entities. Consider, for a moment, the heuristics of decomposition and localisation, which Bechtel and Richardson (2010) associate with the mechanistic framework. Since mechanistic explanations describe systems that are highly integrated, moving any one part can make a big difference to the operation of the whole system. Also, in the case of tightly integrated systems, the particular structure of a component can be essential to explaining how the target phenomenon is produced. For example, the structure of the Venus flytraps trap leaf is essential for understanding how they snap quickly. An explanation or a description that omits facts about spatial organisation is one in which the heuristic of localisation is unlikely to be useful. In such a case, however, it appears that the heuristic of decomposition can still be useful. When a heterogeneous environment is represented in a description of an interacting population, however, the heuristic of localisation can again come into play as spatially local features or sets of features can make a difference to the activities of agents. This is

another sense in which spatially implemented ABMs really do possess the sorts of features expected by the mechanistic framework and also demonstrates that, if uniformity of parts decreases the explanatory value of mechanistic information, then heterogeneity in the environment can increase it once more. This is the case both with Lake Titicaca and the effects of its shape on the distribution of political power in Griffin and Stanish's model and is the case with distribution of habitat for the functional connectivity of the Marten populations in Day et al.'s model.

Before I conclude this chapter, let me briefly say something about Jones' slime mould model. Earlier, I noted that this was a case where an ABM was used despite the target—that is, the slime mould *Physarum polycephalum*—not being a population. Instead, it is a material of interwoven strands of biological fibre. There are no units of Gel/Sol interaction as there are in the model. This is a case where the model does not act like a mechanistic description or explanation. Although it focuses on spatially arranged parts and temporally arranged activities, these parts and their activities exist only as idealisations in the model. Unlike a model of a population, where virtual agents can be mapped onto to real ones, there are no real units of Gel/Sol interaction represented by Jones' agents. Still, there is a clear benefit to cases like these where simple rules can be given to a population of identical components to produce global patterns of interest to the modeller. Although it is beyond the scope of this chapter to go into any depth about this benefit, I suspect that Bechtel and Richardson's argument that one way of tackling complexity is to decompose a system applies in the case of these idealised representations of materials. By decomposing a material into a hypothetical population of units and providing those units with simple rules that are able to reproduce larger patterns, *some* progress has been made towards decomplexifying and demystifying the operation of these systems.

4.7 Conclusion

In this chapter, I have examined the representational capacities of ABMs. Many agent-based modellers have claimed that the value of ABMs comes from their capacity to represent the mechanisms underlying population phenomena and that, previously, population phenomena could only be

represented at the population level. Similarly, mechanist philosophers have also claimed that ABMs could represent mechanisms, indicating that there is something of philosophical substance to the modellers' claims. Other philosophers, however, have suggested that mechanistic explanation is suited to cases where systems are functionally integrated and that populations are better approached with an alternative explanatory framework. This left us with a puzzle: are ABMs mechanistic representations of population phenomena or not?

Ultimately, my response to this question was that ABMs are representations of population phenomena that share many features of mechanistic descriptions. Like mechanistic descriptions, ABMs represent the component parts of a system and their activities. This alone distinguishes them from other representations of population phenomena, which typically approach populations from higher and more abstract levels of organisation. When ABMs are spatially implemented, I have argued that they include all the features of mechanistic descriptions and that the heterogeneity of parts found in paradigmatic mechanistic descriptions and explanations is replaced by heterogeneity in the environment, which means that the mechanistic strategies of decomposition and localisation remain useful and that information about spatial and structural specifics remains explanatorily valuable.

While this chapter has focused on the representational capacities of ABMs, arguing that they are unique among models of populations to the extent that can generate population phenomena from the bottom up, the next two chapters will examine the epistemic weaknesses of ABMs. If these weaknesses are not addressed, then the value of the unique representational capacities explored in this chapter are reduced at best and completely undermined at worst.

5

The epistemology of agent-based models

In this chapter, I examine the relationships between model descriptions and model structures, comparing equation-based models to agent-based models, arguing that using model descriptions that can be manipulated by digital computers leads to uncertainty regarding this relationship. In the case of many equation-based models, digital implementation requires numerically approximating continuous mathematics, leaving some uncertainty about whether the structure investigated through simulation is relevantly similar to the structure which the modellers desire and intend to investigate. Agent-based models may appear to avoid this uncertainty because they typically do not involve the discretisation of continuous mathematics. However, as I argue, uncertainty remains regarding whether their descriptions specify the structures intended. This uncertainty contributes to more general epistemic problems faced by agent-based modelling, which arise from a proliferation of models built in different programming languages and where the exact structures that are investigated in a study can be difficult or impossible to discern from the information provided in publications. One step that can be taken toward addressing these problems is to adopt and enforce standards of model communication in journals that publish agent-based model studies, though this is no easy fix.

5.1 Introduction

As argued in the previous chapter, agent-based models (ABMs) are an important tool for representing interacting and heterogeneous populations. Moreover, they do this from a perspective not offered by equation-based models, which represent these systems at the level of the population rather than the at

the level of its members. While Chapter 4 presented an optimistic argument for the epistemology of ABM, focusing on their unique representational and explanatory capacities, this chapter will examine the epistemic weaknesses of ABM.

Among agent-based modellers, it is common knowledge that ABM has a history of under-delivering on its promise to dramatically change the sciences that use them. With respect to the promises, papers with titles like "Agent-based modelling: A revolution?" (Bankes, 2002) argue that the representational capacities described in the previous chapter make ABMs revolutionary given the historical constraints that limited modellers to representing populations as homogeneous entities. With respect to under-delivering, multiple papers published over the last two decades have described the methodological challenges facing ABMs (An et al., 2020; Angus & Hassani-Mahmooei, 2015; Crooks, Castle, & Batty, 2008; Galán et al., 2009; Hamill, 2010; Macy & Willer, 2002; O'Sullivan et al., 2016; Richiardi, 2017). In their recent paper, Steven Manson and colleagues (2020), note that, disappointingly, many of problems raised in earlier papers have persisted and remain unresolved.

One persistent problem is a lack of proper communication of model detail. Three quick examples will suffice. Dawn Parker and colleagues (2003), discussing ABMs of land use and cover change, state that poor practices of model communication and rates of model replication give ABMs the "image of pseudoscience," encouraging the sharing of model source code as an antidote, emphasising the importance of a "common language through which model mechanisms can be communicated" (p. 331). Similarly, Flaminio Squazzoni (2010) argues that, after about 15 years of agent-based modelling in the social sciences, its failure to produce the dramatic changes some hoped for is in part explained by the lack of a "common methodological standard on how to build, describe, analyze, evaluate, and replicate ABM" (p. 220). In their recent editorial paper, Li An and colleagues (2020, para. 3.2) state that model transparency and reusability remain "one of the bottleneck problems of the ABM community," which limits the usefulness of ABM and undermines the insights that might be drawn exclusively from ABMs given their unique representational capacities. This chapter focuses on the strand of epistemic challenge facing ABMs described by these authors.

To explore the epistemology of ABM and the persistent problem of poor model communication and the associated lack of common methodology, I will begin by contrasting ABMs with equation-based models.[45] An argument made by physicist and sometimes-philosopher Fritz Rohrlich in an early (1990) paper in the philosophy of computer simulation argued that ABMs might have an epistemic edge over equation-based models because they were not based on continuous mathematics which must be discretised to be investigated with computer simulation.[46]

Given this need for approximation in equation-based modelling, he argued, modellers could never be sure that their simulation investigated the abstract structure they desired rather than some other structure. Rohrlich labelled this uncertainty the "model dilemma." You might be sceptical that the model dilemma really is an epistemic bogeyman. Afterall, there is a lot of mathematical machinery devoted to numerical approximation and estimating the extent of errors in these approximations. As I will argue in this chapter, there are still many steps involved in translating desired relationships into a form that can be investigated through simulation even in the absence of approximated continuous mathematics. These many steps, coupled with the fact that ABMs lack any kind of common language like well-established mathematical equations, result in ABM facing its own version of the model

[45] Immediately, and as an anonymous referee pointed out to me, it must be stated that not all equation-based models *are* based on continuous mathematics. The discussion ahead, then, proceeds with the caveat that the scope of Rohrlich's argument is constrained to equation-based models where continuous mathematics is being numerically approximated. If you find this caveat unbearable, skip Section 5.2 and ignore some of the set up in Section 3.2.

[46] That paper contained several suggestions that have remained central to the philosophy of simulation, including the suggestion that computer simulation may be a sui generis scientific methodology sitting somewhere between modelling and experimentation, a matter discussed in section 1.4. I will be leaving any analysis of these grander claims aside. Here, I wish only to consider Rohrlich's claims about epistemic differences between ABMs and the more frequently discussed equation-based models.

dilemma that contributes to challenges around communicating model structure.[47] If you are sceptical about the impact of the original model dilemma, then please judge these positive arguments about ABM's and their epistemic problems on their own merits.

By beginning with Rohrlich's suggestion, I wish to show that the specification relationship holding between model description and model structure is epistemically important, especially in the context of computer simulation where translucent or opaque structures and descriptions are commonplace. As we will see, seeking a method to make the specification relationship as transparent as possible is a task that modellers must devote time and energy toward, with community infrastructure playing a key role in coordinating their efforts. Although motivated by the persistent epistemic problems facing ABMs, my analysis uncovers general lessons for the epistemology of modelling and computer simulation. Before continuing, though, I have one final caveat. As my arguments will show, there does remain some important differences between the case of ABMs and equation-based models. In particular, the problem in the case of ABMs appears to be a practical problem that could be solved, while the problem in the case of discretising continuous mathematics is an principled problem that, if it can be solved, it will be solved with a very different approach. I will discuss these differences at the end of the chapter. However, I will note that none of these particulars

[47] Note that Rohrlich explicitly discussed cellular automata rather than the more general ABMs. Cellular automata are a subset of particularly simple ABMs. In cellular automata, agents are cells of a two-dimensional lattice, which are described in totality by their location in two-dimensional space and a further binary variable (on or off or alive or dead, as in Conway's Game of Life, for example). Finally, agents have a transition rule. I take Rohrlich's focus on cellular automata to be a function of what was common given the technological limitations of the time rather than a principled decision based on the special properties of cellular automata not shared by ABMs more generally. Consequently, I will be applying Rohrlich's claims to ABMs rather than just cellular automata. Moreover, my key aims in this chapter are to demonstrate the epistemic difficulties associated with describing and communicating ABMs. So, if my argument fails as a refutation of Rohrlich's specific claims, then so be it.

detract from the main lesson of this chapter for the epistemology of modelling and simulation, which is the importance of the specification relationship in the epistemology of computer simulation.

Here is the plan for the chapter. In section 5.2, I will describe the model dilemma that simulations of equation-based models face: there is uncertainty regarding whether the simulations investigate the structures desired by the modellers and described by the original equations or whether they investigate some relevantly dissimilar structure. In section 5.3, I will argue that an analogous uncertainty can be found in the case of ABMs despite the absence of approximated continuous mathematics. I will also argue that the epistemic situation is much worse in the case of ABM science because (1) modellers do not have the benefit of a standardised and common language in which to describe their "conceptual models," which are the initial sketches of a model indicating which relationships the modeller would like to investigate in a formal model, and (2) agent-based modellers likewise lack a standardised and common language to describe the implemented, formal models. This anarchic state of ABM specification hampers efforts to understand, compare, and replicate ABMs, which is essential for producing valuable results. In section 5.4, I will describe a solution, which amounts to adopting a range of standardised approaches to model documentation and communication which make the conceptual model as transparent as possible and indicate how it should be implemented computationally. In section 5.5, I will close the chapter by suggesting how the lessons of this chapter generalise beyond the context of ABM.

5.2 What is the model dilemma?

Approximation of physical laws has been commonplace since Newton and Leibniz independently developed calculus and differential equations became physicists' language of choice. When a few of these equations are coupled together to create a more realistic model of a complex target, as when Bjerknes coupled existing laws from fluid dynamics to describe atmospheric flow, the mathematical object described cannot be analysed. That is to say, the equations cannot be solved; they are mathematically intractable. As we saw in chapter 2, modellers can manage this problem by using

approximation techniques such as, in the case of Earth System Models, the use of a finite difference grid.

Rohrlich argues that approximation creates what he calls the *model dilemma*: "any disagreement between the model and the empirical evidence… can either be due to the model qua approximation to the theory, or it can be due to the theory of which the model is an approximation" (Rohrlich, 1990, p. 511). Eric Winsberg also identifies the model dilemma in the context of simulation models and puts it in terms of the Duhem problem, sometimes known as the Duhem-Quine Problem (Winsberg, 1999, 2010): "The Duhem problem is that when a prediction fails to match observed data, we do not know whether to blame the theory we are testing or an auxiliary hypothesis. Similarly, when a computational model fails to account for real data, we do not know whether to blame the underlying model or blame the modelling assumptions used to transform the underlying model into a computationally tractable algorithm" (2010, p. 24).[48]

To better understand Rohrlich's model dilemma and to see whether ABMs avoid it, let's continue looking at Winsberg's work. Winsberg argues that a version of the Duhem-Quine problem of testing holism is found within the inseparability of verification and validation in the case of equation-based models. *Verification*, at least in the context of equation-based models, is a matter of establishing that an algorithm is a good approximation of well-established physical laws. *Validation*, on the other hand, is a matter of establishing that a mathematical structure is sufficiently similar to a target system for that model's purpose. So, when approximation is involved, there are multiple similarity relations in play: one between the original model and the approximation, and another between these models and

[48] Note that I have a different view of models to Winsberg. Although I make a distinction between "the underlying model"—that is, the relationships of interest to the modeller and specified by continuous mathematical equations—and "a computationally tractable algorithm"—that is, the formal language specifying a set of relationships to be investigated with computer simulation—I put this distinction in different terms. For me, these are just two different abstract structures specified with different descriptions. The relevant question, then, is whether the second structure is relevantly similar to the first.

the target system. For the modelling to succeed, both the similarity between original and approximation (verification) and between the model and the world (validation) must hold to a suitable degree and in the relevant respects. Winsberg argues that verification and validation, though separate in principle, are typically run together in practice. Figure 5.1 is a diagram that will help me to explain this problem and why it matters, and to later assess Rohrlich's claim that ABMs avoid the problem.

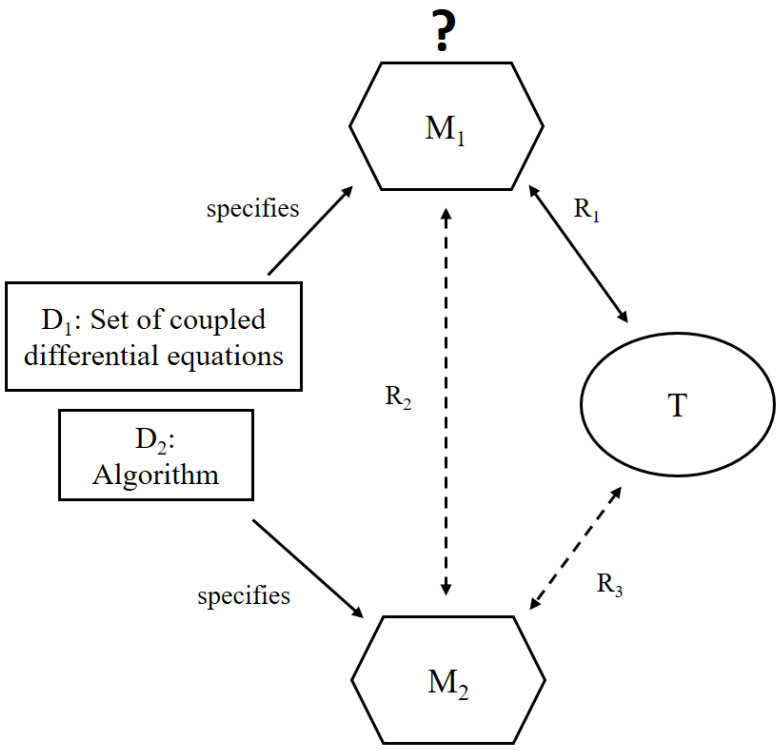

Figure 5.1 The model dilemma conceptualised with a similarity view of models.

Figure 5.1 shows the model dilemma as I conceive of it within the similarity view of modelling described in Chapter 1. A description D_1, written with a set of coupled differential equations, specifies a model system M_1. A question mark hangs over M_1, however, because the model is mathematically intractable, so cannot be analysed; modellers can point to the desired model system with their equations but cannot investigate it because the equations are intractable. This is unfortunate because we have reasons to think that M_1 stands in a good resemblance relation R_1 to the target system *T*. One such reason might be that the components used to make D_1 are known (due to a history of rigorous testing) to describe models resembling, for example, fluid flow, and *T* is just a complex

instance of fluid flow. This is the case with Bjerknes' governing equations representing atmospheric and oceanic flow. The equations are thought to do a good job of representing these systems because the component pieces are known to represent fluid dynamics well and the atmosphere and ocean can be treated as fluids.

Since D_1 is intractable, learning something about T by investigating a model requires using a new description: D_2. D_2 is a set of instructions for a digital computer which specify a model that remains complex but is also tractable—that is, its formal description can be manipulated as a means to investigate the abstract structure it specifies. D_2 specifies an abstract structure M_2, which stands in two resemblance relations, R_2 and R_3, with M_2 and T respectively. These relations are represented with broken lines because the relations are uncertain. Given that the modeller is not really interested in M_2 but M_1, they must take actions to ensure that M_2 and M_1 are relevantly similar.[49] Indeed, it would be ideal if they were the same abstract structure, in which case R_2 would certainly be satisfied. If M_2 was shown to resemble M_1 in the relevant way—that is, to be the same trajectory (or a very similar trajectory) through multidimensional state space—then we would have good reason to think that R_3 would be satisfied. This is because we have good reason to think that M_1 is relevantly similar to T because it is specified with D_1, which is built out of components which are known to describe models resembling T-like systems. At this point, you might think that R3 is the crucial model-target relationship, with M1 (and its R2 relationship) serving only as a guide to the construction and investigation of M2. This is a worry to which I will return below. For now, let's get back to verification, validation, and their inseparability.

[49] As am examiner helpfully noted, numerical analysis is an entire branch of mathematics devoted to determining the similarity between difference and differential equations. Consequently, modellers often have reliable mathematical methods to deploy in their actions. Undoubtedly, numerical analysis has progressed in the last three decades, leaving Rohrlich's original model dilemma far less epistemically damaging than it may have originally seemed. If this is so, the dilemma described in Section 5.3 still stands on its own as a worthwhile topic for analysis.

Establishing resemblance R_2—that is, turning R_2 from a broken line into a solid line—is a matter of verification. When performing verification, the modeller is assessing whether their new description D_2 specifies the structure they wish it to specify. That is, they are assessing whether M_2 is as close to M_1 as it can be. For example, they may know that the mathematical relationships in M_1 embody the laws of fluid dynamics and may now be assessing how well the relationships in M_2 embody the same laws. To verify that their code specifies the desired mathematical relationships, simulation modellers use a range of techniques that fall into two categories: static testing and dynamic testing (Fairley, 1976; Sargent, 2013). Static testing is called such because it does not involve running the program. Instead, it involves systematically and rigorously reviewing the program's code and structure, sometimes producing proofs if possible. Dynamic testing, on the other hand, requires running the simulation program under different conditions and analysing its evolution and results to determine whether it is behaving as it ought to. To determine whether the program is behaving as it should, modellers sometimes compare the evolution of the model to data of other similar systems—either target systems that the model is supposed to represent or other model systems.

Establishing resemblance R_3 is a matter of validation. The modeller is assessing whether M_2 is a good representation of T given their purposes. Validation techniques typically involve simulating M_2 under many different conditions, including those that match the historical conditions of T, and seeing how well M_2 recreates the patterns exemplified by T.

As Winsberg argues and the above description indicates, validation techniques are sometimes used when performing dynamical verification. For example, within climate forecasting, a simulation model will be adjusted according to how well or poorly it recreates historical data and performs on different benchmarks. These sorts of techniques must be used because, as you will recall, the shape of the mathematical structure M_1, described by the well-established equations D_1, cannot be investigated because the coupled equations are intractable. To put it another way, we cannot be sure that the mathematical structure described by the algorithms is the same as that described by the equations because we will never be able to solve the analytically intractable equations and make a direct comparison between the two.

Now, why would this matter? If we have used validation techniques in assessing our D_2 and M_2, then is the resemblance between M_2 and M_1 important? After all, the resemblance was important because M_1 was relevantly similar to *T*, but if we have used validation techniques in our verification, then haven't we already established a resemblance between M_2 and *T*? Indeed, aren't we in a better position now because we have established this resemblance directly? While, at first, this appears to be a reassuring way of dissolving any concerns, some problems remain. One worry is that the data about *T* that modellers use in the dynamical verification could be biased, leading to a model that is relevantly similar to *T* in appearance only. Hang on, if the data can't be trusted to be representative of *T*, then why think that M_1 would be free from any biases impacting M_2? To see the worry, think again of climate models, especially since these are the sorts of models Winsberg has in mind during his argument. Since M_1—think Bjerknes' governing equations—is based on well-established physical principles, the similarity between M_1 and *T*, where *T* is the circulating atmosphere and oceans, is not assumed on the basis of data about the atmosphere and the ocean. Rather, these laws have been shown, through rigorous experimentation, to describe circulating fluids, and the atmosphere and oceans act as circulating fluids; therefore, the relationships of M_1 apply to *T*. To reiterate: M_2 may resemble *T* in appearance alone due to unrepresentative data about *T* used in dynamical verification, while completely different data, and a completely different chain of inference, is used to establish the similarity between M_1 and *T*. In climate science, this is a source of uncertainty: we know M_1 resembles *T*, but we do not know if M_2 really does after all.

Suppose, now, that we are also uncertain of how well M_1 resembles *T*, so R_1 would also be represented with a dashed line. In building D_1, we might not have used all or only the best bits of our theoretical knowledge about complex *T* systems. In this scenario, any dissimilarity between M_2 with *T* revealed when comparing their behaviour may be caused by a failure of any number of the resemblance relations. For example, M_2 may be verified, so R_2 established, but M_1 may still not be a validated model, meaning that R_1 and are R_3 are not satisfied. Alternatively, R_1 might be satisfied, but M_2 may not resemble M_1, meaning that R_2 is not satisfied and R_3 may also be unsatisfied as a result.

As this discussion shows, the model dilemma makes it very difficult to place blame or praise on one part of the web of inference over another. It would be good news for agent-based modellers if ABMs did not face these difficulties in virtue of the absence of numerically approximated continuous mathematics. As I will argue in the next section, despite this absence, ABMs face uncertainties that are very similar or even worse.

5.3 A dilemma of their own

Rohrlich argues that ABMs avoid the model dilemma because they do not involve the approximation of intractable equations (Rohrlich, 1990, p. 511). If our assumptions or hypotheses are originally formalised in discrete, logical form, rather than continuous mathematics, then, according to Rohrlich, we can implement it on a computer without recourse to approximation.[50] Returning to figure 5.1, Rohrlich's argument suggests that, in the case of ABM, we avoid the need for the lower part of the diagram and D_2, M_2, R_2, and R_3. This is because the lower part of that diagram concerns a redescription of the original model. In turn, this should reduce the uncertainty when making inferences about where to place praise and blame. This is because D_1*, our discrete form description of a hypothetical ABM, is tractable, unlike the intractable mathematical equations of D_1. So, we can investigate M_1* and, if we find a match or mismatch between M_1* and T*, we can place the praise or blame on M_1*. In this section, I will argue that ABMs face their own version of the model dilemma, so are not epistemically superior to equation-based simulations for this reason.[51] In fact, practices around describing models are an epistemic weakness in ABM science.

[50] It is worth noting that the absence of continuous mathematics is not unique to ABM. Here, my focus is on whether the distinction between continuous and discrete models (unique or not) is epistemically relevant in the present case. Still, it is useful to remember this deeper problem for Rohrlich's original argument.

[51] Note that I do not expect ABMs to be epistemically superior to equation-based simulations for any other reason nor the other way around. At least not in general. As the last chapter described, ABMs can have an edge of equation-based models when representing some systems, such as heterogeneous interacting populations, for

Here is a sketch of my argument. Although agent-based modellers may not begin with some set of equations that they must approximate, they also do not begin by simply writing lines of code in the language of the program they ultimately wish to use. Instead, they start by describing what is sometimes called a *conceptual model* (Sargent, 2013; Wilensky & Rand, 2015). To put this in Winsberg's terms from the beginning of the previous section, there is still the distinction between an "underlying model" (the original abstract relationships not specified with computer code) and the model specified with "a computationally tractable algorithm" created with additional modelling assumptions. I will argue that the relationship between an implemented ABM and the initial conceptual model is sufficiently similar to the relationship between a numerical approximation and initial equations in the case of equation-based simulation to generate a version of the model dilemma which is faced by agent-based modellers. Indeed agent-based model-builders recognise the presence of this version of the model dilemma (Galán et al., 2009, paras. 1.3-1.4): "a prerequisite to understanding a simulation is to make sure that there is no significant disparity between what we *think* the computer code is doing and what is actually doing… This problem is particularly acute in the case of agent-based simulation."

To make this argument, let me start with some more details about conceptual models. Uri Wilensky is an agent-based modeller responsible for creating NetLogo, a user-friendly program for building agent-based models. In his text-book with William Rand, the two describe conceptual models as flow-chart diagrams or pieces of pseudocode which map out the parts of the simulation and their relationships (Wilensky & Rand, 2015, pp. 314–315). Rather than thinking of conceptual models as the flow-charts or pseudocode themselves, I think of them as relationships which can be specified with these descriptions. Box 5.1, for example, shows a slice of pseudocode Jones' includes in his book describing the sensory behaviour of the particles comprising his virtual plasmodium (see section 1.2.3

---

some purposes, but the superiority or inferiority of a model must be judged on a case-by-case basis rather than in virtue of the descriptive or structural form chosen.

for more details), where “F,” “FL,” and “FR” refer to the particles’ front, front left, and front right sensors, and “RA” stands for a random amount (Jones, 2015, p. 38):

```
[Sensory stage]

- Sample chemoattractant map values
- If (F > FL) & (F > FR)
        - Continue facing same direction
- Else if (F < FL) & (F < FR)
        - Rotate by RA towards larger of FL and FR
- Else if (FL < FR)
        - Rotate right by RA
- Else if (FR < FL)
        - Rotate left by RA
- Else
        - Continue facing same direction
```

Box 5.1 A pseudocode description of the virtual slime’s sensing algorithm.

In Chapter 1, I mentioned that Godfrey-Smith has argued for a view of model structures according to which mathematical models are imagined, fictional worlds, rather than mathematical objects (Godfrey-Smith, 2006, 2009b). Although I disagree that this is the most helpful way of conceptualising of mathematical model structures, I do think Godfrey-Smith is right to point out that imagined scenarios play an important role in model-based science. One way in which they can prove useful is in assisting the construction of a conceptual model. Although I am not adopting a models-as-fictions view of conceptual models, my bet is that, if any scientific models were fictional scenarios, it would be conceptual models. This is because the function of conceptual models is to provide a guide to the main entities and relationships that will be investigated with a formal model, but conceptual models need not be mathematically or rigorously manipulable themselves. This manipulability is a crucial feature of formal models, whether they be described by equations or lines of code but is inessential and typically lacking in a conceptual model.

Agent-based modellers start with a conceptual model for multiple reasons. For one, it is better to have a plan of attack before jumping into writing lines of code—just as a philosopher will probably start a paper with an outline of their argument, a programmer will start with an outline of their

program. More specifically, the relationships of interest should be mapped out before efforts are made to implement these relationships algorithmically and in the programming language of choice. More interestingly, the person or group that desires the simulation results and specifies which functions the program should compute might not be the person or group who performs the coding, delegating this job to a specialist (Humphreys, 2009, p. 624). In such cases, it is very important that these people agree about just what the simulation model needs to include such that it can test the intended hypotheses.

Alternatively, a modeller may wish to replicate someone else's results. In that case, they must start by reconstructing a conceptual model to see which relationships are included in the target model that need to be replicated with their own program. Consider Jones' virtual plasmodium. Suppose that I wanted to replicate Jones' work in a different language, like NetLogo. I might start by identifying the elements of Jones' simulation and the functions they perform, writing up chunks of pseudocode like the one above. The simulation would start with, for example, a command like "populate area with a random distribution of particles with density *D*." Such a command produces my initial conditions of a random distribution of agents with some density. Once I had mapped out the procedure my model would follow, I could then set to work coding the model in NetLogo's programming language.

Although flow charts and pseudocode may describe the intended relationships more transparently, they cannot be used to manipulate and investigate the model. Just as before, a new structure must be described in a language that allows for investigation through computer simulation—that is, the conceptual model be implemented computationally. Figure 5.2 shows the model dilemma as it applies to ABMs. In this figure, M1* is the conceptual model; it is the entities and relationships the modeller would like to investigate given their hypothesis, the principles and dependencies they are exploring, and so on. M2* is also a set of entities and relationships the modellers wish to investigate but described in a form that actually permits the investigation of the model because the description is a computer program. For the analysis to produce insight into the modellers' hypotheses about their target or the principles they wish to explore, M2* must contain the entities and relationships of M1*.

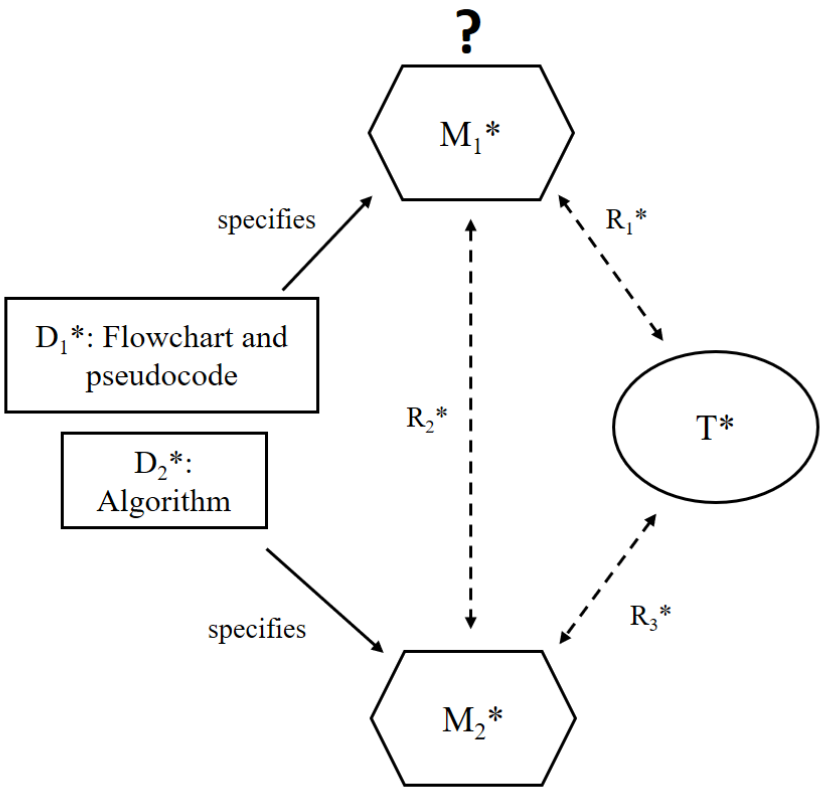

Figure 5.2 A model dilemma for agent-based models.

Walking through figure 5.2 demonstrates that agent-based modellers must still perform verification and validation to ensure that M2* embodies the mathematical relationships desired and, if the model is to be compared to the world, that those relationships are sufficiently similar to relationships found in the target. Establishing resemblance R_2* is a matter of verification. Wilensky and Rand suggest building ABMs function by function and module by module, testing that each block of code performs the intended task as it is supposed to. Still, they warn, components that individually perform as intended may not perform as intended when coupled together. Establishing R_1* and R_3* is a matter of validation. Remember that we cannot validate M_1* directly because it is not described in a form that allows it to be systematically investigated. M_2*, on the other hand, can be validated because it can be manipulated through computer simulation, which allows its behaviour to be compared with data about T*. However, the modeller will only really be able to assess their original hypotheses or assumptions when they are confident that M_2* is sufficiently similar to M_1* to embody and exemplify

the same principles and assumptions. Once again, we see Rohrlich's model dilemma: we do not know where to place the praise or blame when things go wrong or right.

Things may be even worse in the case of ABMs than they were in the case of equation-based models. There, we might doubt the importance of the relationship between the implemented model structure and the conceptual model structure because the implemented model structure is getting tested directly against empirical data about the target system. Although I argued that uncertainty regarding the quality of our empirical data may cause modellers to rely, to some degree, on the history of success of the relationships in the conceptual model, in the case of ABMs, we often do not even have recourse to direct testing against the world. A major epistemic problem all of its own facing ABM, to which we will return in the next chapter, is that of the connection between ABMs and empirical data and how to validate ABMs (O'Sullivan et al., 2016; Wallentin, 2017). Given this problem, we should be less confident about relying on validation to ensure the quality of our model than in the case of, say, climate science. Setting this particular problem aside until Chapter 6, however, we might be especially concerned about the relationship between the implemented model and the original relationships of interest in the conceptual model because many ABMs are made with the express purpose of theorising, reasoning through assumptions, exploring possibilities, and so on, rather than predicting the behaviour of real systems (Edmonds et al., 2019; Epstein, 2008). In that case, the purpose of the simulation model *just is* to investigate the relationships of the conceptual model. So, if the simulation model fails to resemble the relationships of the conceptual model, it fails to investigate them and fails to serve its primary purpose.[52]

To make matters worse, the situation depicted in figure 5.2 is a simplification, omitting other steps in the model-building process. Agent-based modellers José Galán and colleagues (2009), in a

[52] Here too we might think that the prospects are dimmer in the case of ABM than in the case of equation-based models. Since equation-based models involve the implementation of conceptual models that are themselves specific formally, mathematical machinery exists for comparing the behaviour of these models with the computer simulation models. In the case of ABM, there is no such formal apparatus.

discussion of ABM errors and artefacts, distinguish between at least three different roles in the generation of ABMs. The “thematician” is responsible for designing the conceptual model, the “modeller” turns the conceptual model into an algorithm, and the “computer scientist” is responsible for implementing those algorithms in a programming language such that the model structure can actually be investigated through simulation. Of course, some or all of these roles may be played by the same individual, but they need not be. This means that inference from simulation result to information about a principle is a multistep chain:

> The analysis of the results of the computer model leads to conclusions on the behaviour of the *computer scientist*'s model and, to the extent that the *computer scientist*'s model is a valid approximation of the *modeller*'s formal model, these conclusions also apply to the *modeller*'s formal model. Again, to the extent that the formal model is a legitimate particularisation of the non-formal model created by the *thematician*, the conclusions obtained for the *modeller*'s formal model can be interpreted in the terms used by the non-formal model. Furthermore, if the *modeller*'s formal model is representative of the *thematician*'s model, then there is scope for making general statements on the behaviour of the *thematician*'s model. Finally, if the *thematician*'s model is satisfactorily capturing social reality, then the knowledge inferred in the whole process can be meaningfully applied to the target system. (Galán et al., 2009, para. 2.22)

One very salient instance in which these roles are not played by the same person is in the context of model replication, where one researcher or modelling group attempts to recreate the work of another. According to modellers Edmonds and Hales, (2003 section 12.2, c.f. Grimm et al 2006), model replication carries particularly large epistemic weight in the context of ABM because formal analysis, which can be performed only on mathematically tractable equations, is not possible in these cases: “If we are able to trust the simulations we use, we must independently replicate them.”

Wilensky and Rand (2007) have argued that, with a few notable exceptions, there has been alarmingly little replication of ABMs. Using multiple models is a common method for increasing

confidence about a model's result or otherwise making better inference from models. This may come in the form of model ensembles of opportunity or more systematic robustness analysis (see Chapter 3). In order for a modelling community to perform robustness analysis, they must be able to replicate one another's models, especially since, for robustness analysis to succeed, aspects of the model must remain the same in order to show that the core dependencies exemplified by the model are robust to variations of the other construction assumptions and idealisations.

This lack of replication is not simply due to a lack of motivation among agent-based modellers; rather, modellers face practical challenges because of the ABM model dilemma, which discourage them from performing replication. ABM replication is hindered because "published descriptions of [ABMs] are often hard to read, incomplete, ambiguous, and therefore less accessible" (V. Grimm et al., 2006, p. 116). For example, Brian Heath and colleagues' (2009) survey of 279 studies published between 1998 and 2008 found that "68 unique software packages or programming languages" were used across these articles (para. 4.2) and, alarmingly, "104 articles (37.3%) did not provide any details on what package or programming language was used to construct and execute the simulation" (para. 3.2). Similarly, in their (2021) review of 32 obesity ABMs from 2013 to 2019, Philippe Giabanelli and colleagues found that only two studies shared their source code. Their review also found that, at least according to the measures they were using, there was little to no reason to think that model studies were getting any better even if there were more studies over time.

This lack of consistency is not isolated to the formal languages used to specify model structures for investigation through computer simulation. Rather, there is a similar amount of variation regarding how the model and its results are described for the purposes of communicating the model study and its results to other scientists. Simon Angus and Behrooz Hassani-Mahmooei (2015) conducted a systematic analysis of papers published in the *Journal of Artificial Societies and Social Simulation*, a journal specialising in ABM, between 2001 and 2012 to produce a review of the ABM communication methods used in the journal over that time. They found that all manner of communication modes had been used, such as tables, flow-charts, pseudo-code, and so on, and that there was no convergence on a standardised form of communication. Perhaps more damning than the

variation is the kind of information included in these pieces of communication. As they state, “60% of these 'results' figures didn't clearly indicate the parameters used to generate them” (para. 5.8).

Here again we have a point of difference between ABMs and equation-based models. The way in which I have been arguing might suggest that ABMs alone have conceptual models, but this is not true. There are conceptual models for equation-based models as well. So, one might think the situation is still worse in the case of equation-based models because a conceptual model is still the basis of a mathematical model, which is then approximated numerically with computer code. Although there is uncertainty regarding how well equation-based simulations approximate the relationships of their conceptual models, there is at least a common formal language for describing these relationships. That is, the translation process from conceptual model to mathematical model is more transparent due to a common mathematical language which is used to formalise the kinds of relationships included in these conceptual models. This common language is sorely absent in the case of ABMs. In the case of ABMs, there is no common language for either conceptual models or formal, computationally implemented models. Consequently, model communication and the replication that communication facilitates is far more difficult in the case of ABMs.

According to Wilensky and Rand (2007), faithful ABM replication frequently requires direct consultation with the original modeller because the published description of the conceptual model is simply too sparse to permit proper reconstruction of the modelling decisions made. Discussing Wilensky’s replication of Axelrod’s iterated prisoner dilemma,[53] Wilensky and Rand provide some

[53] In Axelrod’s iterated prisoner dilemma model (Axelrod, 1984), a population of agents is divided into pairs. The pairs then play an iterated prisoner’s dilemma for 200 rounds. In this way, the model is a tournament for the different agents. The agents, in this case, play different strategies in the iterated prisoner’s dilemma. Probably the most famous strategy, and certainly the one that won the first tournament, is that of tit-for-tat, where an agent opens by playing cooperate in the first round and then mimicking the other player’s previous move for every subsequent round. The winning strategy of the tournament was determined by which, across these pairs, was able to achieve the highest score against all other strategies and itself.

examples of simple implementation differences that can make a difference to the model's overall behaviour: order of interactions between agents (serial or simultaneous), order of events (immigrate, interact, birth, death vs. immigrate, birth, death, interact),[54] order of agent action (shuffled vs. ordered: "For instance, if there was a concentration of altruistic agents in the upper left corner of the world and they always got a chance to reproduce before any of the agents in the bottom right, then they would not have to compete with the agents in the bottom right for space and would come to dominate the population") (Wilensky & Rand, 2007, para. 5.19). The comments from these modellers presented above not only demonstrate that the connection between a conceptual model informally described and a computationally implemented model is often unclear, meaning the ABM model dilemma is a real phenomenon facing ABM practitioners, but also that it is a problem with real epistemic consequences.

Let's take stock. In this chapter, I have so far argued that the uncertainty regarding how well a chosen formal model description specifies the relationships of the desired model structure can be found both in the context of equation-based models, which involve the numerical approximation of continuous equations, and agent-based models, which typically do not. This runs counter to the view that ABMs would mostly escape the uncertainty precisely due to the lack of numerical approximation. The uncertainty regarding ABM specification is a problem for two reasons. The first is that models are built to investigate relationships that are of interest to the modeller for any number of reasons and the purpose of modelling is undermined if the simulations do not actually provide any information about these relationships. The second is that replication is crucial to ensuring the epistemic value of models, but replication efforts are undermined if modellers only share and have access to rough natural language descriptions of the relationships investigated by their models. Proper model replication requires both very clear descriptions of the relationships of the original conceptual model as well as the formal description used to investigate these relationships in the model study being replicated. While the best-case scenario might be direct consultation with the modeller or modellers, this is

[54] Note that in this second sequence, it is not the case that all steps happen to the same agents. For example, any given agent is not immigrating before they are born or interacting after they die. Rather, existing agents immigrate, reproduce, those flagged for death will die, and the remainder interact.

simply infeasible as widespread solution. Instead, in section 5.4, I will describe how a standardised model documentation framework can be used to improve the communication of model structures within ABM science and support epistemically important practices like model replication.

5.4 Improving the epistemic state of ABM

The epistemic value of ABMs has been harmed by a laissez-faire attitude toward model communication, which has the consequence of obscuring the actual mathematical structures that are being investigated within model studies, limiting the ability for other modellers within the community to interpret, replicate, or extend the work of their peers. In this section, I will outline a general strategy or set of strategies that ABM science should adopt to improve its epistemic value. In brief, this is to adopt a standardised framework for detailed model description, along with the practice of storing source code in online repositories that can be accessed with ease.

The need for standardisation within ABM science has been recognised by multiple modellers. Angus and Hassani-Mahmooei, whose (2015) review was described above, argue that the community should impose and enforce stricter regulations on communication, most likely in the form of submission guidelines. All going well, these would increase the transparency of ABMs and increase their credibility in the eyes of editors and referees of general scientific journals beyond those that specialise in publishing ABM research. Likewise, Heath and colleagues, whose (2009) review found a proliferation of formal descriptive languages and many papers omitting details on the software packages used, argue that standardised model documentation is required such that models can be replicated irrespective of the programming language chosen to build the original model, which is especially important if the original model is built in a proprietary software, so cannot be freely shared. Most simply, Willensky and Rand (2007) suggest that lines of pseudo-code should be published alongside the formal model, expecting that modellers will converge upon a standardised language to better facilitate model replication. For example, Candelaria Sansores and colleagues (2005) propose a tool-kit-independent visual language for specifying simulation models for communication and

replication. This language essentially uses box-arrow, flowchart-style diagrams to show the different relationships of the different entities within the ABM.

Modellers have long been calling for standards and standardisation in ABM science but, for the most part, there is little evidence that any are being implemented. There is some hope, however, that this persistent problem will be addressed. One framework for documenting ABMs that has seemingly stood the test of time and appears to be becoming more widespread is the Overview, Design concepts and Details (ODD) protocol, developed by Volker Grimm and colleagues over the last 15 years (2006, 2010, 2020). In section 5.4.1, I will describe this framework before suggesting some of the ways in which it can assist in section 5.4.2. Of course, you may worry that uniformity comes with a cost. I will also address this worry at the end of section 5.4.1.

5.4.1 An ODD solution

The ODD is essentially a set of seven elements to guide modellers on how to provide complete documentation of their models. The first three elements, falling into the *overview* part of the schema, are descriptions of: (1) the purpose of the model and the patterns it tries to recreate; (2) the entities, state variables, and spatiotemporal scale of the models; and (3) a process overview and scheduling. The fourth element, *design concepts*, contains a list of 11 design concepts, some of which may not appear in a model: (i) basic principles; (ii) emergence; (iii) adaptation; (iv) objectives; (v) learning; (vi) prediction; (vii) sensing; (viii) interaction; (ix) stochasticity; (x) collectives; and (xi) observation. An ODD for Jones' virtual slime, for example, would omit concept (v) because the units in Jones' virtual plasmodium model do not learn at all. Furthermore, if there are design concepts not already on this list, they can be added at the end of the design concepts section of the ODD. The last three elements fall under the *details* category: (5) initialisation; (6) input data; and (7) submodels.

The ODD protocol has gotten traction in modelling circles, with the initial (2006) paper having received 2822 citations and the (2010) update having received 2331 citations according to Google Scholar as of mid-May 2021. Moreover, the Journal for Artificial Societies and Social Science

(JASSS), a journal where many ABMs are published, has suggested that modellers use the framework along with the CoMSES catalogue—an online ABM repository—to provide the details of their model and better facilitate model understanding and the possibility for replication within the community. The ODD has also been extended by other modellers to make it more widely applicable beyond its original domain of application. For example, the ODD+D adds an element for describing "decisions," and was intended by Birgit Muller and colleagues (2013) to be a means of facilitating the application of the framework to economics and social science, where the decision rules that agents use are more varied than in the case of ecology, where the framework was originally formulated. Furthermore, and arguing that ODD and ODD+D have no specific place for modellers to reflect upon and record the relationship between their model and data, Ahmed Laatabi and colleagues (2018) introduce the ODD+2D, which adds another D for "data." This demonstrates that the core organisational idea behind the ODD is a good one, requiring refinement and patching to address its limitations rather than complete overhaul or replacement.

To provide a clearer picture of an ODD, box 5.2 contains a summary ODD for Jones' model that I have attempted to reverse engineer. Grimm and colleagues recommend that modellers write a full ODD for their model and have it evaluated by a peer before writing a summarised ODD, which can be included in the body of the paper. The full ODD can then be attached to the paper as an appendix. It should be noted that Grimm and colleagues emphasise the need for specificity in the description of all the elements, including the model's purpose, stating that, if the description of the model's purpose is vague, it will be difficult to evaluate the model relative to that purpose. Since a complete ODD can span many pages, Grimm and colleagues suggest including a summary ODD in the body of the paper. Grimm and colleagues provide templates for both ODDs and summary ODDs, suggesting that, as much as possible, modellers use the prompt phrases provided in these templates to assist in the standardisation of ODDs. In box 5.2, these phrases are left in bold and the ODD's keywords are italicised, as directed by Grimm and colleagues in the summary ODD template. Where possible, I have quoted Jones verbatim, showing how an ODD and ODD summary can help

communicate a model's description and construal that might have otherwise been distributed throughout a publication.

The overall *purpose* of Jones' model is to "reproduce the complex biological and computational behaviour of *P Polycephalum* which is manifested in its complex foraging, growth, pattern formation, network adaptation, oscillatory transport, and amoeboid movement" and to illustrate that "simple component parts and local microscopic interactions [can] generate the complex macroscopic behaviours in an emergent, bottom-up, manner" (p. 4). **To consider his model realistic enough for its purpose, Jones uses the following patterns**, which are "those observed in *Physarum*," including growth and adaptation of plasmodial networks: "in an environment with high nutrient concentration," a radial growth and formation of "a uniform sheet-like plasmodium behind the growth front," followed by symmetry breaking "as protoplasm is transported to different regions of the plasmodium, forming a protoplasmic network which becomes sparse over time" (61). This includes how network adaptation is affected by environmental hazards, like repellents and light, and the ability to distinguish between "the closest, strongest, and largest sources of nutrients in uniform and noisy environments" (p. 86). Another pattern the virtual plasmodium aims to replicate is Physarum's spontaneous "oscillatory patterns, pattern transitions and synchronisation behaviour" (p. 269). I will not repeat all the quantitative descriptions of these patterns, but these quantitative specifications come from precise measurements of the Physarum in experimental conditions that can be replicated virtually. An example of the precise numerical specifications of these patterns are that nodes in a Physarum network have an average connectivity of about three (nodes are typically connected to three other nodes).

The model includes the following *entities*: A diffusive lattice, "represented by a discrete two-dimensional floating-point array"; a discrete lattice, which is coupled and isomorphic to the diffusive lattice and which stores the locations of the agent particles; and the particles of Gel/Sol interaction. In an ideal ODD summary, the details of these entities, such as their parameters and variables, should be provided in a table, which would include the variable's or parameter's label,

the kind of values it can take, which values are typical or chosen in the model, and the interpretation of the variable or parameter by the modeller. For example, the gel/sol units in Jones' model have the variables GROW? and SHRINK?, which are binary and take a value according to the growth and shrinkage rules described in section 1.2.3. The interpretation of these variable is that they represent whether the conditions are favourable for this area of the slime mould to grow or whether the plasmodium material is likely to die off or retract in this area. I have elected not to provide a full table of all the variables and parameters in this summary ODD not out of laziness—although I think one or two examples would suffice for our present illustrative purposes anyway—but because I could not provide a complete table. While Jones does describe many of the parameters and variables of his model in publications, I could not be sure I had described them all without looking at the model's code. Unfortunately, the model, along with all the other supplementary material of Jones' (2015) book, is no longer available online (in 2020) as it was only a couple of years ago. This only speaks to the problems that follow from lacking a centralised and standardised repository of models or the inclusion of thorough and rigorous descriptions accompanying every publication. Indeed, if I could easily reverse engineer an ODD for Jones' model, it would suggest that the protocol has limited value as all the relevant information would already be in the publication.

The most important *processes* of the model, which are repeated every *time step,* are that the units move, the units are tagged for shrinkage or growth, and that the chemo-attractant diffuses.

The most important *design concepts* of the model are *emergence*, *sensing*, and *interaction*.

(ii) *Emergence:* The virtual slime is based on emergence, with Jones aiming to demonstrate how adaptive network formation, along with other aspects of Physarum's flexible behaviour, can emerge from the collective interactions of agents that act as the plasmodium material.

(vii) *Sensing*: Agents sense chemoattractant in the diffusive lattice as well as environmental hazards like light. This sensing is performed using front-facing sensors, the angles and reach of which can be altered but are set as 45 degrees left and right of centre and to a distance of 50 pixels as a

default. This is done to represent the abilities of the plasmodium material to sense these same environmental features—that is, environmental attractors and repellents. Agents sense without noise, which is an idealisation.

(viii) *Interaction*: Agents interact but only via the diffusive lattice. When agents move, they secrete chemoattractant into the diffusive lattice, which affects the behaviour of other agents as their heading is determined by concentrations of chemoattractant.

Box 5.2 A summary ODD for Jones' virtual slime.

Before I turn to a discussion of the ways in which the ODD framework can assist in tempering ABM's problem of anarchic communication practices, let me briefly respond to anyone who might be concerned about the negative consequences of standards and standardisation.

Implementing standards and standardisation raises its own set of questions: have we implemented the right ones? Consider, for a moment, the case of significance testing in psychology where a p-value of 0.05 has long been the standard for a result considered good enough to publish. Naturally, this caused perverse incentives for p-hacking, where data collection is manipulated to increase the chances of finding statistically significant results without going so far as to use false data (Wiggins & Chrisopherson, 2019). For example, a researcher might collect information on a number of different variables to increase their chances of finding an effect but only report the variable that produced an effect, or a researcher might collect data only to the point where they find a significant result, and then collect no more. Even in the absence of these practices, however, finding a statistically significant effect does not mean the effect is real, as there may be other methodological problems with the experiment's design, or the experimenter may simply have gotten lucky. Afterall, if the standard is a p-value of 0.05, then about 1 in 20 statistically significant results should be a false positive.

The standards of psychology's null hypothesis significance testing are very much unlike the standards for which I am advocating in this section. Most obviously, the p-value of 0.05 is a kind of success criterion, while the ODD framework for model documentation model is not. Rather, this framework is much more akin to the standards that have been suggested in response to psychology's

replication crisis, which was connected to its standard of significance testing. For example, researchers in that field have been encouraged to improve their practices around the transparent communication of experimental methods to better facilitate replication by other research groups, as well as clearly stating all and only those variables for which they tested, and sharing the data and code used to analyse the data (Wiggins & Chrisopherson, 2019). This is very much like calling for modellers to share their code and provide clear documentation.

Indeed, there are further similarities between the two cases with the implementation of a badge system to signify which papers and studies comply with these best practices. In psychology, there are three badges: one for those who have preregistered the details of their experiment before they have collected any data; one for those who have made their data publicly available, including in a protected access repository if the data includes sensitive information; and one for those who make accessible the digitally shareable materials and methods required to perform the experiment.[55] Where journals such as *Psychological Science* use these badges, sharing has increased, demonstrating that they do have a positive effect on improving transparency (Kidwell et al., 2016). In the modelling context, and not specific to ABM, the "open code" badge is awarded to those who have made their source code "publicly available in a searchable, open access, trusted digital repository," which should be accompanied by detailed documentation and metadata required to understand, replicate, and execute the source code (CoMSES Network, 2020).

Still, the question remains: if we advocate for a standardised model documentation framework, have we chosen the right one? The simple answer to this is that any reasonable framework is better than none. In part, this is because adopting and using a framework can reveal where the framework falls short and needs revision. Adopting a standardised framework does not mean that the framework cannot change, and it is much easier to judge what is required from a framework when one

[55] You can find more details about these badges on *Psychological Science*'s website (https://www.psychologicalscience.org/publications/badges), and a list of Journals that use these badges can be found on the Centre for Open Science's website (https://www.cos.io/initiatives/badges).

is being used than it is from the armchair. Recall that the ODD was noted to have some limitations when applied outside its original domain of ecology, which were met with extensions to the framework in the form of the ODD+D (Müller et al., 2013) and ODD+2D (Laatabi et al., 2018). Regardless of the standardised protocols used, they will require refinement. But only through this process of refining a framework can they be incrementally improved and tailored to suit their various applications. In the meantime, modellers will have a much better chance of understanding one another's work, replicating it, and extending it. I will now turn to a discussion of how the ODD and other steps towards standardisation can be used to mitigate the persistent problem of ABM opacity.

5.4.2 The benefits of ODD

The ODD framework and its extensions assist in improving the epistemic state of ABM in at least two ways: (1) increasing model transparency; and (2) supporting model replication, comparison, and robustness analysis. In the remainder of this section, I will examine these benefits and the ways in which the ODD supports their realisation.

First, and most obviously, using a framework like the ODD reduces model opacity by providing detailed documentation of the model's structure along with the motivations and intentions driving the representational choices behind the model. Importantly, this detailed description is neutral with respect to any particular programming language which could be used to formally specify the structure of an ABM. This is important because, as mentioned in section 5.3, there is a proliferation of programming languages used to specify the structures of ABMs. While an abundance of programming languages may not, by itself, be a problem, it is a problem to the extent that modellers cannot be expected to know every programming language their colleagues use, so cannot be expected to understand every model from the source code alone. Moreover, it would be nearly impossible to convince every modeller to use the same programming language or to enforce such uniformity in journals. On the other hand, using a standardised documentation framework in the same natural language ensures the communication of model structures and intended interpretations in a shared

language and is already recommended by journals like the *Journal of Artificial Societies and Social Simulation*.

The ODD would be a useful framework for reducing opacity even in the absence of a proliferation of programming languages. Even if all ABMs were written in the same language, and the source code for all models was made freely available, the ODD would be useful because ABMs are used in many different disciplines, such as ecology, economics, and epidemiology. This diversity has an effect similar to a diversity of programming languages. Different disciplines may have their own technical terms, communication standards, and theoretical principles, which can all contribute to difficulties understanding and interpreting models.[56] Indeed, one of the express purposes of the ODD is to facilitate better cross-disciplinary interaction between modellers (V. Grimm et al., 2020). Furthermore, computer code must be interpreted and can easily be misinterpreted. Providing a clear description of the procedures and principles implemented in the code, then, can assist in accurate model interpretation. Indeed, given the benefits of this function of model documentation, in their latest update to the framework, Grimm and a host of other authors (2020) recommend including a rationale section at the end of each element in the ODD dedicated to the motivation behind the representational choices made.

[56] How much do modellers in one discipline require or desire to interpret and understand the work of those in another discipline? If the answer is "not at all," then opacity between disciplines is hardly a problem. However, I do not think this is the right answer. This is because, despite their different subjects, agent-based modellers share a methodology: ABM. As an anonymous reviewer pointed out, different disciplines may very well share an interest in the same abstract structure, though with an alternative interpretation. Consequently, there may be lessons regarding the methodology or insights into manipulating these abstract structures that are valuable across disciplinary divides. Folks may well be scratching their heads in one discipline because of a problem that has already been solved in another. While opacity between disciplines is far less concerning than opacity within disciplines, it still carries a cost relative to the goal of improving the methodology of ABM (V. Grimm et al., 2020; Lorscheid, Berger, Grimm, & Meyer, 2019).

Does the ODD, a standardised method for documenting ABMs, along with providing source code in an online repository like the CoMSES catalogue, completely solve the problems of model opacity? Perhaps not. As Wilensky and Rand (2007) note, nothing beats a conversation with the model builder to provide a clear picture about the operation of their model (see also De Regt, 2014; Kaiser, 2009). However, in the absence of this, providing detailed documentation of the intended model structure and the relationship between the model description and this structure in a standardised format like the ODD, as well as providing the model description source code wherever possible, will go a long way to increasing transparency within ABM science.

Let me now put this benefit of the ODD in terms of verification, validation, and the framework shown in figure 5.2. With respect to verification and validation, the ODD most clearly assists with verification. Recall that verification is a matter of ensuring that a formal model behaves as intended relative to the construction assumptions used to build it, allowing modellers to make judgements and inferences about these assumptions, though perhaps not the world. Questions regarding the relationship between the model and the world are questions about validation. The ODD assists primarily with verification because it documents both the conceptual model and the implemented model, stating in natural language the kinds of relationships and design principles that ought to be included in the model, and stating, at least an abstract level, how the formal model goes about implementing these relationships and ideas. When the ODD is complemented with the source code, the relationship between the conceptual model and implemented model is made even clearer. So, by providing the details of the conceptual model and the ways in which they are implemented in the formal model, the ODD provides direction on what the model ought to be doing and how to assess what it would look like if it were verified. That is, the ODD does not ensure models are verified, but makes the conditions under which the model would be verified more explicit.

Although the primary value of the ODD is the insight it provides regarding the internal structure of the model, there are still great benefits to be had regarding validation with an ODD. For a start, by stating the target patterns against which the model is to be judged, the ODD suggests how one could validate a model and what level of similarity between model and target is required for the model

to count as validated given its purpose. Furthermore, validation efforts can be improved if the ODD+2D extension of the framework is used because it is dedicated toward explicitly describing the connection between model and data.

Although the ODD makes the relationship between conceptual model and implemented model far clearer, it still does not entirely remove the problem of where to place the blame when things go right or wrong. Recall that the model dilemma Rohrlich attributed to equation-based models was that a model's performance could be attributed to the original equations and relationships underlying the model or their computational implementation. In the case of ABMs, the ODD assists in assessing whether the computational implementation is a good one and when it is a poor one. Hence, it may be able to help settle the question of where to place the blame; if we know the ABM is a good faithful implementation of the underlying relationships, then the ABM may be to blame. However, there are very many different ways in which a modeller may computationally implement a formal model. So, if a model produces some result, the modeller may remain unsure about whether to attribute that result to the hypotheses and assumptions used to construct the conceptual model or those introduced only when implementing the model. One way in which we might go about assessing this is to vary computational implementation while keeping the conceptual model fixed, which would be a form of robustness analysis.[57] As I will argue next, the ODD can help here too.

The second way in which a standardised model documentation framework like the ODD can assist with ABM's anarchic state (and with the model dilemma for ABMs) is by facilitating model replication, comparison, and robustness analysis. The way in which the ODD assists with replication should be clear: the ODD includes detailed documentation of the patterns the model aims to generate, along with the entities, procedures, and parameters used to generate its focal patterns. Quite simply, it is a comprehensive natural language description of the model which includes details that have been,

[57] Specifically, it would be an instance of what Weisberg and Reisman call representational robustness analysis (Weisberg & Reisman, 2008).

for the most part, excluded from the model presentations provided in journal articles, and was made precisely to improve the chances of model replication.[58]

The ODD and its extensions improve that capacity for model comparison because they provide scaffolding by decomposing models into elements. This decomposition allows any reviewer to search for similarities and differences across models by looking at each of these elements. For example, Uta Berger and colleagues (2008) construct ODDs for three existing models of mangrove ecosystems, allowing them to quickly chart the features the models shared and those they do not; two of the models represent resources as heterogeneously distributed, one does not, and so on. As papers like these suggest, while having ODDs already accompanying model studies would assist in model comparison, there is still an advantage to constructing ODDs for pre-existing models because the organisational framework of the ODD provides dimensions along which models can be compared. A reviewer can look at the design concept of *interaction*, for example, and find that only two models explicitly represent competition for light while the other does not distinguish between the different resources for which trees compete (Berger et al., 2008).

Without a detailed documentation framework, the dimensions along which models should be compared must be devised anew each time someone undertakes model comparison. While these new frameworks may be superior to the ODD, the chances of this occurring decrease as the ABM community continues to refine the ODD framework. Moreover, if the ODD is already in full use, then anyone looking to compare models will already know where to look for details and will not need to reconstruct ODDs for the models they are comparing themselves. Furthermore, while it may be possible to reverse-engineer ODDs in some cases, given the scant details often available in publications, this will frequently be impossible. Although the ODD is valuable insofar as it provides a framework for comparing models, its value is vastly increased if everyone is already using the

[58] As earlier remarks indicate, where the ODD falls short of facilitating a comprehensive documentation of model details, it has been extended with the ODD+D (Müller et al., 2013) and ODD+2D (Laatabi et al., 2018) and through successive updates (V. Grimm et al., 2010, 2020).

framework. And, of course, having the details of models decomposed according to the ODD framework does not stop anyone from using alternative dimensions along which to compare models. Rather, it simply makes information regarding the same elements[59] easier to find across different models, making comparison along almost any dimensions easier.

As argued in Chapter 3, robustness analysis is epistemically important, providing information about the causal relationships holding within and across model structures. Understood as the use of multiple, closely connected models with systematic variations, robustness analysis is assisted by a standardised framework like the ODD in at least two ways. The first is that, if one modeller is to create a variation of another's model, they must first be able to replicate the model before varying it. If they cannot replicate the original model, then they have not shown that they have varied it. The ODD assists with robustness analysis, then, because it assists with replication, and replication is a crucial step in robustness analysis. Importantly, the ODD's programming language neutrality provides an advantage relative to providing the source code alone. That is, it increases the ease with which models can be replicated faithfully in different languages. This is itself a form of robustness analysis, demonstrating that a result does not depend on some quirk of the formal language used to implement the model (Weisberg, 2013).[60]

[59] By this, I mean the elements of the ODD, such as the overview element, the design concepts element, or the details element.

[60] To be clear, the connection between varying programming languages and robustness analysis and robust theorems is as follows. Robustness analysis requires a common causal core. In this case, it is the relationships (at least some of them) comprising the conceptual model and which can be implemented with different languages. By using these different languages, the modeller can help identify the causal core, which is the collection of elements required to produce the focal result, and the scope and limits of the dependency between causal core and result. For example, if they reproduce every relationship in a different language and find the same result, then they know the result is robust at least with respect to these two languages. This is an instance of "representational robustness," mentioned in Chapter 3 (Weisberg & Reisman, 2008). They may then continue to

The second way in which that the ODD can assist with robustness analysis is related to its impact on model comparison. As mentioned above, the ODD assists in model comparison by decomposing a model into its elements, including the entities included, the rules they follow, the parameters used, the actions taken each time-step, and so on. As I've argued, this makes model structures more transparent and, consequently, easier to compare. However, this decomposition by elements also suggests ways in which the model can be varied. A modeller engaged in robustness analysis can examine the 'entities' element of the ODD and consider what might be missing, or they may look at the 'decision' element of the ODD+D and revise the decision procedure governing an agent's behaviour.

As this discussion demonstrates, implementing a standardised model documentation framework like the ODD and its extensions, along with providing source code in accessible repositories, would go a long way to improving the epistemic state of ABM. It would increase model transparency and support the practices of model replication, comparison, and robustness analysis. These three practices are key methods for understanding why models produce the results they do.

5.5 Conclusion

In this chapter, I have argued that ABMs do not face the same version of the model dilemma as equation-based simulations, like ESMs, because their construction does not involve the discretisation of continuous mathematics. Consequently, there is no concern for how well the original structures specified by continuous mathematical equations resemble those specified by the discretised versions included in the computer simulation model. However, ABMs still face their own version of the model dilemma. This is because implementing the relationships comprising a conceptual model with a

vary the model by adjusting different structural assumptions to assess which of the replicated relationships belong in the causal core and which do not and what belongs in the *ceteris paribus* conditions.

programming language still involves creating a new description and a new structure even if it may not involve a switch from continuous to discrete mathematics.

The uncertainty about the extent to which the model system specified by the new description resembles the system specified by the old description remains despite the absence of some of the problems specific to numerical approximation, such as round off and truncation errors. Indeed, the persistence of a model dilemma for ABMs and its epistemic importance has been recognised by the modelling community, and is related to broader problems regarding the difficulty of knowing what model structures are investigated in model studies on the basis of the loose and incomplete natural language descriptions found in publications. The opacity of ABM model structures has been identified as a source of challenges for model interpretation, communication, and replication, which are real and serious epistemic problems for the modelling community.

I have argued that the best way to improve model opacity is to introduce standardised protocols for documenting ABMs. A thorough documentation framework like the ODD is programming language neutral and comprehensively details the abstract structure being investigated. When accompanied by the source code, this documentation allows other modellers to better evaluate the extent to which the source code specifies the intended structure, especially if the ODD is used to explain the rationale behind modelling choices. The ODD can also be used to replicate models and perform robustness analysis, concentrating investigation onto the same or related model structures. It can also be used for model comparison, connecting perhaps previously disconnected models.

It is true, however, that the model dilemma in the context of ABMs is a practical problem with a different flavour to the model dilemma in the case of equation-based models. Accordingly, the way in which these problems are addressed differs. While uncertainty regarding numerical approximation might be addressed with mathematical techniques directed toward evaluating approximations, the model dilemma in the case of ABMs is addressed with practices aimed at improving the accessibility of conceptual models, formal models, and the descriptions thereof. However, there are further

differences that complicate the ABM case, such as the absence of a common language for describing conceptual models like the mathematics that informs equation-based models.

In making this argument, this chapter contributes to the philosophy of computer simulation by showing that the persistence of the model dilemma beyond the domain of numerical approximation is a key part of the epistemology of computer simulation, at least in the context of ABM and highly complex dynamical models like ESMs. This is that, compared with models that are simple enough to be investigated without resorting to simulation, there are further difficulties determining whether model descriptions specify the mathematical relationships desired or whether they instead specify nearby relationships. This, in turn, complicates the chain of inference from simulation results back to the assumptions thought to underly the model's behaviour. This uncertainty is absent in the context of much simpler models which have formal descriptions that are transparent to modellers within that field.

How general is this feature? Will all computer simulation models face these problems? Although I am cautious of making overly general statements about modelling, I can describe the two main drivers of the uncertainty, and we can see that many computer simulation models will be affected by these drivers. The first driver is that the modeller wishes to specify the same relationships while changing model description. This change always introduces the possibility that the modeller will, with their additional model descriptions, specify a model structure that is relevantly different from that specified by the first model description. The second driver of the uncertainty is that, when dealing with model structures sufficiently large to require investigation through computer simulation, the complicatedness and complexity of both the description and the model structure mean that modellers have a much harder time assessing (a) whether their model description specifies the relationships intended and (b) what the features of the mathematical structure they intend to specify are such that they can identify, from the model's behaviour, whether or not this is the structure they wished to specify. To put it another way, dealing with big mathematical objects that must be specified with big model descriptions creates opacity in both.

In the next chapter, I will continue discussing the epistemology of ABMs, focusing on ABMs that are intended to represent specific targets for the purposes of contributing to policy decisions. We will see that, while implementing the practices described in this chapter would assist in this domain, there are deeper problems for ABMs which prohibit them from being as successful in this domain as high-fidelity climate models.

6

The scope and limits of brute force modelling

In this chapter, I examine the practices required to address the epistemic challenges facing highly complex models by comparing high-fidelity climate models with agent-based models. I argue that thoroughly investigating highly complex simulation models requires a proportionally large amount of community coordination. While examples of such community coordination can be found in climate science, they are mostly absent in the context of agent-based modelling. My arguments also reveal that heterogeneity among target systems creates a further problem for complex models. In addition to hampering the generalisability of complex and detailed models, target heterogeneity divides resources, prohibiting research communities from establishing the practices that are required to ensure that complex models perform as well as possible. Consequently, fields characterised by target heterogeneity are likely to have less success with high fidelity modelling than fields focused on a singular target or small set of similar targets.

6.1 Introduction

In Chapter 2, I argued that increasing the size of a model's structure might permit a more comprehensive representation of a target but that this came at a cost. "Brute force" models—models which aim toward a very high degree of realism or detail—face three epistemic challenges as described by Levins (1966): brute force models are data-hungry; brute force models are computationally expensive; and brute force models can be difficult to understand. As Levins argued, a consequence of this is that there will always be a role for models with simpler structures that do not face these epistemic challenges. However, Levins' argument, and my arguments of Chapters 2 and 3, leave us with an unanswered question: What should modellers do if they wish to pursue the brute force approach?

In this chapter, I will argue that addressing the challenges facing brute force models can be overcome only with a concerted effort by a scientific and epistemic community to establish the right sorts of institutions and infrastructure that (1) enable that collection and maintenance of empirical data and model data, (2) assess the impact of approximations and idealisations used in the face of computational limitations, and (3) conduct analyses of model behaviour to identify the internal causal relationships. To make my argument, I will show that climate science has much of the institutions and infrastructure in place to mitigate the challenges of brute force modelling. This case study demonstrates both that mitigating the challenges of brute forcing is possible, at least to an extent, but also that doing so comes at a very large cost in terms of human and financial resources. Indeed, as I will argue, climate science might be uniquely positioned to receive the resources required to perform brute force modelling at its most epistemically respectable.

I will compare climate science with target-directed modelling in ABM science, where the same challenges of brute force modelling can be found. However, I will argue that, due to the nature of the targets that ABMs are used to represent, the prospects for mitigating these challenges in the way climate science does are dim. Where ABMs represent human decision making, in particular, or decision making distributed throughout institutions like firms or governments, it is very difficult to collect accurate data and, in many cases, it's not even clear what good data would look like. Likewise, improving the comprehensibility of models by incrementally improving upon a relatively small set of models which are thoroughly investigated by a large community of researchers is unlikely to prove effective in ABM science because ABMs are used to represent heterogeneous targets, which divide research communities and multiply representations. While there may be some scope for developing a set of basic building blocks that could be used as the basis of different models, which could allow for the continued investigation of these basic sub-components, it is not clear just how far these sub-components would generalise. The lesson to draw from this is that we should expect the brute force approach to only succeed primarily in cases where scientists are concerned with a shared target and there are great gains to be made from being able to predict and manage its behaviour. This reveals that

there are further costs to target heterogeneity beyond the modelling trade offs that Levins identified. Target heterogeneity reduces the capacity for scientists to pool resources and tackle the challenges that follow from building highly complex and complicated simulation models.

Here is the plan for the chapter. In section 6.2, I will briefly recap the main features of high-fidelity climate models and the ways in which they face the three epistemic challenges Levins described. In section 6.3, I will describe the Coupled Model Intercomparison Project (CMIP), which is an impressive institution that coordinates the evaluation of state-of-the-art climate models and argue that this institution has contributed to solutions to the challenges facing brute force models. In section 6.4, I will contrast this case to that of ABM science. I argue that the targets of ABMs possess features which prohibit the use of many of climate science's successful strategies for dealing with the challenges of brute force modelling. In section 6.5, I will conclude that some targets are less amenable to a successful brute force approach and are better off settling for lower or middling degrees of realism and comprehensiveness.

6.2 Brute force climate models

In this section, I will provide brief refreshers on two things. First, in section 6.2.1, I summarise the main features of Earth system models (ESMs). Second, in section 6.2.2, I provide a summary of brute force modelling and the ways in which ESMs are brute force models. If these details are fresh in your mind and you do not need a refresher, feel free to skip ahead to section 6.3.

6.2.1 A brief refresher on ESMs

The main components of ESMs and the similar global circulation models (GCMs) are two modules comprising the *dynamical core*. These two modules represent the global atmosphere and the global oceans respectively as circulating fluids. For each of these components, space is represented with a three-dimensional grid, where values for regions of space—values for variables like temperature and

air pressure—are updated according to numerical approximations of the laws of fluid dynamics. These laws themselves are far older than climate models, with the set of seven non-linear differential equations having been originally articulated in their coupled form by Vilhelm Bjerknes in 1904 (Bjerknes, 1904; Edwards, 2010; Washington & Parkinson, 2005, p. 49). However, these equations cannot be integrated simultaneously, so cannot be used without numerical approximation.

In addition to circulating atmosphere and ocean components, these models are composed of other modules and parameterisation schemes that represent other elements of the climate system. For example, additional modules are used to represent vegetation, atmospheric chemistry, and aerosols, which are all outside the scope of fluid dynamics. Parameterisation schemes are used to represent processes, such as cloud formation, that cannot be resolved bottom-up at the resolution of the model. These components are collectively known as the *model physics*. The relationships encoded within these modules are based on the, sometimes incomplete, empirical knowledge of climate scientists. Although a finer resolution model may represent its target more accurately, resolution comes at the price of computational power and time, constraining model resolution to a scale that requires the parameterisation of sub-grid scale phenomena.

With this summary out of the way, let's turn to what makes ESMs and GCMs instances of brute force modelling.

6..2.2 A brief refresher on brute force modelling

In Chapter 2, I argued that high-fidelity climate models are paradigmatic instances of brute force modelling. "Brute force" modelling is a term that Levins coined in his (1966) response to criticisms from systems ecologists like Kenneth Watt and C. S. Holling, who Levins derided as performing "FORTRAN ecology" (Levins, 1968a). Unlike Levins and his colleagues, the systems ecologists optimistically saw new computational technology as an opportunity to create highly detailed models that could include the vast number of mathematical relationships that must be omitted from analytically tractable mathematical models. Brute force modelling, then, refers to a strategy where

modellers, leveraging the power of digital computers to perform many calculations very quickly, build models with structures that are as comprehensive as the limits of computational power and empirical knowledge will allow.

Returning to the key components of the simple similarity view, described in detail in Chapter 1, and depicted in figure 6.1, we can characterise brute force modelling as follows. Brute force models have large structures comprised of many mathematical relationships, often with complex interactions between them. These large structures allow the representation of as many details about the target system as possible. Given the size of the model structures, brute force models have correspondingly large and complicated descriptions. These descriptions can only feasibly be manipulated, and consequently their model structures investigated, through computer simulation. Brute force models have narrow targets. This follows naturally from the amount of detail included in the model structure, which excludes the possibility of accurately representing the many other possible targets that do not embody similar mathematical relationships. Finally, brute force models aim for as high a degree of both structural and dynamical similarity as possible. They aim to represent as many of the target's processes as relationships in the model structure as possible and aim to represent the changes in those processes over time and in response to interventions as accurately as possible. This allows brute force models to perform their function, which is usually something like prediction and forecasting, providing information about how a system might behave in the future and in response to different interventions. This allows users of brute force model results, such as fishery managers in the case of systems ecology or policymakers in the case of climate models, to either manage the target system more effectively or adapt to the expected behaviour of the system.

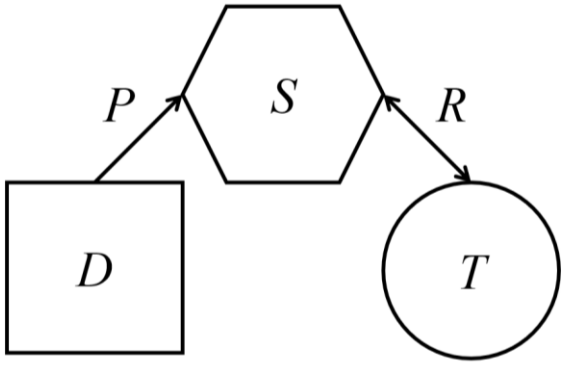

Figure 6.1 The key relationships of the simple similarity view. A model **d**escription *D*, stands in the s**p**ecification relationship *P* with a model **s**tructure *S*, which stands in the **r**esemblance relationships *R* with a **t**arget system *T*.

As Levins identified, three epistemic challenges follow from these characteristics of brute modelling. First, brute force models are data hungry. This is because they both have large structures comprised of many mathematical relationships and because they aim toward a high degree of structural and dynamical similarity between structure and target. As more components and relationships are included in the model structure, more empirical data is required to calibrate the model to ensure a fit between the model and target and to evaluate the performance of the model. Without this data, adding more relationships to the model structure becomes a form of guesswork and can impair the model's ability to provide useful information about the target system and its responses to different interventions.

Second, brute force models are computationally demanding. The model descriptions required to specify the very large structures of brute force models must be manipulated with powerful digital computers. Increasing the size of the structure and, consequently, the description, requires yet more computational power. Computational limits can be a problem for two reasons. First, modellers may be forced to adjust their models for the sole purpose of making them more computationally tractable, replacing the mathematical relationships that they would like to investigate with relationships that are merely good enough given the constraints. This is particularly unfortunate when modellers have prior

reasons for thinking that the original relationships were good representations of the target but are unsure of the resemblance between the replacement relationships and the target. Second, computational limitations can be a problem to the extent that they prohibit modellers from investigating their models as thoroughly as they might otherwise. If computer simulation is required to investigate these vast model structures and simulations are very time expensive, then less investigation is possible in any finite amount of time. This consequence of computational limitation feeds into the last epistemic challenge.

Finally, the size of brute force model structures and descriptions makes them opaque and difficult for modellers to understand. As described in Chapter 3, understanding is a matter of grasping causal relationships and being able to anticipate the behaviour of a system in response to various interventions. When a model has a structure comprised of very many relationships all of which can interact with one another, the behaviour of these structures becomes much more difficult to predict and modellers have a harder time of anticipating how a model system will respond to interventions. This is worsened by the computational limitations just mentioned. First, these limitations lead modellers to alter relationships in the name of added computational efficiency, introducing further uncertainty into their expectations of the system's behaviour. Second, where computational limitations prevent modellers from conducting as many simulations as they otherwise would, they prevent modellers from using the primary means by which they chart the behaviour of a model system and its various components in response to different interventions and conditions.

As I argued in Chapter 2, high-fidelity climate models like ESMs and GCMs are instances of the brute force approach, consisting of large model structures, large model descriptions, and a narrow target. Although these models are not causally complete—if this were a necessary condition of brute force modelling, then no models would be brute force models—the history of their development reveals that greater detail has been included as computational and scientific advances have allowed. This is the best indicator of models aiming toward the goal of completeness given the impossibility of actual causal completeness.

As expected, given the characteristics of brute force models, high-fidelity climate models face the three challenges described above. First, the empirical data sets used to calibrate and evaluate climate models are incomplete and it is very difficult or impossible to gather more empirical data to complete historical data sets even if we can gather more empirical data for future data sets. Second, the structures comprising climate models are the products of computational limitations, including approximations and parameterisation schemes that would be eliminated under ideal circumstances. Likewise, the number of simulations that modellers can perform is fewer than it would be under ideal conditions. Third, climate models are difficult to comprehend and the experiments that are used to investigate their internal causal structures are affected by computational limitations.

With this summary of brute force modelling and ESMs complete, let's now turn to the ways in which the climate science community has attempted to overcome the epistemic challenges facing their high-fidelity models.

6.3 CMIP and the three challenges

How have climate scientists attempted to manage the epistemic challenges and uncertainties that face their high-fidelity models? In Chapter 3, I described the process of ensemble modelling, where model results are analysed and presented collectively rather than individually. At first glance, this might seem like a good way to improve the inferences we make with our models. After all, trends that are stable across the models, or results calculated as averages across the models, are not likely to be biased by the construction assumptions, including approximations and idealisations, of any one model. However, as we saw, these ensembles are ensembles of opportunity, where the models make up a very small sample of possible model space, where the sample is not constructed systematically, and where the sample is likely to be biased because the models are insufficiently independent (W. S. Parker, 2009b, 2011, 2013b, 2018). More importantly for our present discussion, ensemble modelling does not have the power to address all the challenges facing brute force modelling. Indeed, it's not obvious that it's especially well-suited to addressing any of them. Instead, we can find climate science's solution to the

uncertainty of their high-fidelity models if we look at the community infrastructure behind the multi-model ensembles: The Coupled Model Intercomparison Project (CMIP). Below, I will give an overview of CMIP before demonstrating how it addresses the three epistemic challenges in sections 6.3.1-6.3.4.

Established by the World Climate Research Program and its Working Group on Coupled Modelling in 1995, CMIP is one of the primary institutions directed toward minimising uncertainty in high-fidelity climate models (Eyring et al., 2016; Meehl, Boer, Covey, Latif, & Stouffer, 2000; World Climate Research Programme, 2016). The intention behind CMIP is to facilitate comparison between, and collective evaluation of, state-of-the-art climate models. This is done by specifying a set of simulation runs or experiments that must be performed by participating modelling centres, which then submit their data to a central repository, allowing direct comparison of different models relative to the stipulated performance metrics. Note that CMIP is not the only model intercomparison project running in climate science, and my focus on CMIP is not intended to imply that any non-CMIP-endorsed model intercomparison projects are unimportant. Rather, I focus on CMIP because it is a very clear example of a piece of community infrastructure the purpose of which is partly to address the epistemic challenges that accompany highly complex simulation modelling.

Initially, CMIP was primarily concerned with identifying model errors and examining how state-of-the-art climate models produced climate variability. During these first steps, CMIP did not use models to explore the possible consequences of greenhouse gas increases as they do today. Instead, the simulations ran until the virtual climates settled into their own natural equilibria and were then evaluated according to how well they could reproduce observed patterns. This is reflected in the specific purposes of CMIP's first phase, which were: (1) identifying errors in coupled models by comparing their performance to observed data, (2) documenting the effects of correction terms in the flux exchange between ocean and atmosphere, and (3), assessing the ability of models to simulate climate variability (Meehl, Boer, Covey, Latif, & Stouffer, 2000). While the meanings of purpose (1) and (3) are relatively transparent, let me report a quick example of a flux adjustment provided by Gerald Meehl and colleagues (Meehl, Boer, Covey, Latif, & Stouffer, 1997). Recall from Chapters 1

and 2 that the flux coupler is responsible for communicating values between the model's atmosphere and ocean modules. Comparing model results to observations may reveal, for example, that the model systematically overestimates sea surface temperatures due to inaccuracies in its representation of clouds, which would reflect incoming radiation, producing the observed lower temperatures. To compensate for this error without producing a better representation of clouds, adjustment could be made at the point where information about temperatures are transferred from the atmosphere component to the ocean component. Flux adjustments are also an example of the challenge of comprehensibility described in section 6.2. That is, modellers may not know exactly why their models systematically over- or underestimate some value, so they introduce a quick fix. This, in turn, can complicate their understanding of the model's behaviour and the downstream consequences of the quick fix. The initial stage of CMIP, then, can already be seen to address at least one of the challenges facing brute force models.

The scope and purpose of CMIP's first phase can be contrasted with CMIP today. Presently, CMIP coordinates the completion of a host of different simulation experiments,[61] including the different forcing scenarios included in the reports of the Intergovernmental Panel on Climate Change (IPCC). To give some examples of the sorts of forcing scenarios that are used, the fifth assessment report (Stocker, 2014) used the scenarios RCP2.6, RCP4.5, RCP6.0, RCP8.5, where "RCP" stands for "representative concentration pathway." Very roughly, RCP2.6 was the best-case scenario, where CO2 emissions declined from 2020 to zero by 2100. In RCP4.5, emissions decline from 2040. In RCP6.0, they decline from 2080, and in RCP8.5, they rise right through to the end of the 21st Century. This work in projection began in the second phase of CMIP, known as CMIP2, which introduced the first forcing scenario in which atmospheric CO2 is increased at a cumulative rate of 1% annually, leading

[61] All I mean by "experiment" in this context is that a simulation is run or "spun up" until the virtual climate reaches its own equilibrium. Then some condition is imposed on the virtual climate, such as an increase in CO2 emissions. The result is then observed and recorded.

to a doubling of CO2 in about 70 years. CMIP3 added further experiments beyond the 1% increase of CO2 introduced in CMIP2, including the scenarios associated with the IPCC reports.

As mention of CMIP1, 2, and 3 suggests, CMIP proceeds in phases, with CMIP6 being the most recently completed. Each phase provides participants with time to run the simulations required of that phase, to analyse the results, and to publish their conclusions, typically in a special issue covering each aspect of that phase. In terms of the internal structure of CMIP, each phase is broken down into sub-projects addressing areas of interest. In the most recent phase, there were 21 model intercomparison projects dedicated toward aspects like cloud feedback and land-use (Eyring et al., 2016). In total, there were 42 articles (43 if we include the introductory overview article) in CMIP6's special issue in *Geoscientific Model Development*, addressing topics beyond those covered by the 21 model intercomparison projects, such as data management (Balaji et al., 2018; Juckes et al., 2020; Petrie et al., 2020) or the creation of accurate historical data sets (Hoesly et al., 2018; Meinshausen et al., 2017; van Marle et al., 2017).

Running CMIP in phases is useful as it provides an opportunity for the organisation to adapt and update its methodology in the face of changes in models, technology, and empirical knowledge. If new model components are introduced, then new experiments can be added to assess those new components, and if technology permits the removal of an idealisation and approximation, then the associated experiment can likewise be removed. CMIP6 also saw a major revision in its approach and structure, rather than just the experiments performed. Following a period of consultation with modellers and end-users, four main issues were identified that were to be addressed in CMIP6 and beyond. Three of these issues will be particularly relevant to my discussion below regarding the connection between CMIP and the epistemic challenges facing brute force models. These were: (1) the need for standardised model data, allowing comparison of models across CMIP phases and outside of CMIP; (2) the need for guidance to allow smaller modelling groups to contribute to specific aspects of CMIP; and (3) the need for CMIP phases to be more targeted on the particular problems faced across the community (Eyring et al., 2016, pp. 1938–1939).

To summarise, CMIP is a project that has been running for around 25 years. Initially aimed toward assessing and improving high-fidelity climate models, CMIP has coordinated simulation experiments for many years, which provide ensemble projections from the world's high-fidelity climate models. The project runs in phases, and, over time, has become increasingly complex, leading to a restructure in the most recent CMIP phase. As we can see, CMIP has two primary functions, relating to the purposes and challenges of brute force modelling. With respect to the purposes of forecasting and providing insight into how an important target system might behave in response to interventions, the experiments that CMIP coordinates provide results from a range of models, which collectively provide insight into the possible behaviour of future Earth climates. Additionally, CMIP uses model intercomparison and its place as a prominent piece of community infrastructure to address, as best as it can, the three challenges faced by brute force models.

In the remainder of this section, I will present, one by one, my view of the ways in which CMIP attempts to address the challenges that arise when taking the brute force approach.

6.3.1 Addressing data hunger

The first epistemic challenge facing brute force models is data hunger. As described in section 6.2, this follows naturally from the aims of causal completeness and predictive accuracy. If a model is built with the intention of satisfying the aims of causal completeness and recreating the dynamics of a detailed target system, then more empirical data must inform the representation of each additional component and relationship and must be used to assess the model's accuracy. Unfortunately, data does not come for free, and can be difficult or even impossible to collect. While climate science will continue to benefit from, for example, cheaper and more abundant satellites going into the future, which will provide important data that will assist in the improvement of climate models, increases in

funding and technological advances will not enable the collection of valuable data for historical time periods.[62]

Initially, it may seem unclear how CMIP and its operations assist with the problem of data hunger. Afterall, collecting climate data largely appears to be a problem with financial and engineering solutions rather than one with solutions coming from community infrastructure. However, this is because, as I have described the problem of data hunger above and earlier in Chapter 2, it is a problem of data *collection*: scientists must collect empirical data to inform and evaluate their models and that data is difficult or even impossible to collect. While that is certainly one part of the problem, the development of CMIP reveals another, significant part of the problem: the problem of data *management*. This is a problem for which community infrastructure plays a much larger role in the solution.

Turning our attention to the management side of data hunger is beneficial as it was not anticipated by Levins, at least not in his published work on the topic of brute force modelling. The epistemic challenges facing brute force modelling might yet be deeper than the sceptical Levins initially realised. Furthermore, solving the first part of the problem of data hunger by collecting more data serves only to feed the problem of data management. Suppose, for example, that climate scientists had all the funding and technology required to record every datum needed to calibrate and evaluate the most sophisticated climate model possible, including a time machine permitting the construction of comprehensive historical data sets. In this world, researchers would still face the problem of data management, requiring infrastructure to archive the growing mass of data in a form that could easily be accessed and exploited by modellers around the world. This task not only requires housing the data somewhere safe and stable in physical space (remember that data, even data "in the cloud," is ultimately physical) and ensuring that researchers in distant physical places can access that data, it also

[62] As mentioned in Chapter 2, this is not quite true. There are ways of collecting more data about past trends, such as by drilling out more ice cores. However, there are limits regarding the kind of further historical data that can be found which cannot be solved with greater funding, more advanced technology, and new techniques.

requires that the data is sorted into sets with standardised formats so that it can be used for constructing and evaluating models efficiently.

The bad news does not stop there. The management side of data hunger is made more troublesome by the practice of ensemble modelling where the results of many simulation runs are stored for the purposes of model intercomparison. While Levins described high-fidelity models as data-hungry, it might be more accurate to describe them as *data-bulimic*. Not only do they possess a great appetite for data to be used in their construction and evaluation but produce yet more data in the form of their results. Data-bulimia also has the same relationship to data management as data collection does. That is, as the challenges of brute force modelling are addressed—perhaps increased computational power allows more detailed representation of the target, addressing the challenge of computational limitations, or perhaps more modelling centres are able to participate in CMIP's ensemble modelling and intercomparison projects, improving our understanding of the effects of different parameterisation schemes—the burden of data management increases. This is because more models, and bigger models with more relationships, produce more model data.

Data management, a key part of any solution to the challenge of data-hunger, requires institutions or community infrastructure, and CMIP helps coordinate this effort. CMIP plays this role in part because the amount of data they have produced for the CMIP projects has increased as models have become more complex and the number of variables tracked by the CMIP projects has increased. Consequently, an infrastructure of data archival has been established which can store the data for efficient retrieval and proper analysis (Balaji et al., 2018; Petrie et al., 2020). To what extent has the data management problem been solved, and what about the remaining problem of missing data? As the mention of a hypothetical time-machine mentioned a few paragraphs ago suggests, the data management problem is easier to solve than the problem of data collection; no science-fiction technology is needed to archive climate data and organise it in such a way that can be accessed and used as easily as possible by researchers. All that is needed is technology that exists, like computer storage and the internet, along with human efforts to ensure the data is standardised. As the last round of CMIP indicates, researchers are establishing this infrastructure for data management and will only

improve this in the future. The problem of data management can and will be solved (Sun, Di, Cash, & Gaigalas, 2020).

We cannot be quite as optimistic about the problem of data collection, though optimism is still warranted. Although there are constraints regarding what data scientists can possibly collect, researchers devise new methods for producing and improving historical data sets, including reanalysis, which involves the use of models to fill gaps in existing historical data sets. Moreover, when considering the severity of the problem of data collection, we must consider how much data is enough data for our purposes. This depends on the exact questions the models are used to answer, and the precision needed to answer those questions. For example, if an ensemble of models tells us that business as usual emissions will lead to temperature increases of about 4 to 10 degrees by the end of the 21st Century, this range does not allow us to make a precise cost-benefit analysis and to calculate the exact dollar value we should be willing to pay to take action on climate change. However, it is enough to direct social planners to do something rather than nothing. Moreover, the most important empirical data may be for specific causal processes within the climate system, such as the behaviour of clouds or impact of aerosols on the atmosphere rather than fine-grained data about the oceans and atmosphere going into the past two centuries. We should be much more optimistic about collecting this crucial data about current causal processes and fine tuning our representations of climate components than we should about the ability to collect data about the past.

6.3.2 Addressing computational limitations

The second epistemic challenge facing brute force models is that these models are more computationally demanding than simpler ones. To some extent, the problem of computational demand looks less severe today than it did in 1966 when Levins' paper was published simply because of the technological advancements made in the intervening half century. Nevertheless, modellers are still unable to build the models they would under ideal circumstances, and as they seek to represent more physical processes in the wake of computational advances, they butt up against the new limits of

computational power. The primary epistemic consequences of computational limitations are, then, that modellers must introduce approximations and idealisations into their models, which might impact the model's behaviour in unwanted or unexpected ways.

A core part of CMIP is concerned with addressing the first consequence of computational limitation. Since its inception, one of CMIP's purposes has been the assessment of approximations and idealisations included in models, starting with assessing the impact of flux adjustments in the first phase of CMIP. In the most recent phase of CMIP, particular CMIP-endorsed model intercomparison projects were targeted toward, for example, investigating the impact of different parameterisation schemes on model disagreement (Webb et al., 2017).

While the use of intercomparison projects to evaluate representational choice is an important practice in managing the consequences of computational limitations, intercomparison projects would remain required in a world where computational power was no object. In that case, the intercomparison projects would be tasked with evaluating yet more complex representations and more intercomparison projects would most likely be required to evaluate the choices behind the proliferation of model components. To put it another way, computational limitations constrain the representational choices modellers make, but modellers will still be required to make representational choices even if these constraints are removed. Therefore, the sort of assessment that model intercomparison facilitates would persist even if modellers could build and investigate the most comprehensive structures imaginable; just because you can represent every desired process in unprecedented detail, does not mean that you will do so correctly.

Given that model intercomparison plays a crucial epistemic role in assessing the accuracy and adequacy of model structures, it is important to recognise that these projects are impacted by the challenge of computational limitations. This too has been reflected in the actions of CMIP. As mentioned earlier, CMIP6 included a consultation period with modellers and end-users. This revealed that the many experiments and scenarios involved in CMIP were becoming burdensome for some research centres, which did not have the resources to perform all the required simulation runs. Rather

than requiring participants to complete every experiment and contribute models to every ensemble, the model intercomparison projects are now conducted as more autonomous CMIP-endorsed model intercomparison projects, which focus on specific details, like the representation of clouds or the representation of ice sheets. To participate in one of these intercomparison projects, all contributors must complete a small set of common experiments known as the Diagnostic, Evaluation, and Characterisation of Klima experiments (*Klima* is Greek for "climate," with the somewhat obscure word used to achieve the acronym DECK). This consists of only four simulation experiments, plus one historical simulation of, roughly, the past 150 years. Moreover, these are all experiments that modellers are likely to perform already as part of model development, so should not be overly burdensome. Under this new structure, it is possible for a smaller research group to complete these common experiments and participate in a single CMIP-endorsed MIP (model intercomparison project) or a small handful of MIPs, while a larger research group with greater computational resources could complete the common experiments and participate in every CMIP-endorsed MIP.

To summarise, computational limitations create two epistemic challenges for modellers: modellers must make compromises regarding their representational choices, choosing to build and investigate model structures that they would not under ideal circumstances; and modellers must make compromises regarding the extent of the analysis to which they subject their model structures, prioritising some directions of investigation over others. Community infrastructure like CMIP can assist with both. First, CMIP has coordinated multi-model analyses with the purposes of examining the impacts of representational choices, uncovering ways that models can be improved in the future or, at the very least, determining which parts need improvement. Second, CMIP has developed a set of core experiments for demonstrating the basic behaviours of a model system and a new organisation that allows modelling centres to choose how to contribute to model intercomparison projects beyond this core set of experiments.

Model intercomparison, and ways of making participation in model intercomparison possible, is important not only for improving the accuracy of model structures but also for improving the community's understanding of their models, the epistemic challenge to which we turn next.

6.3.3 Addressing comprehensibility

The third epistemic challenge for brute force models is comprehensibility. I will note from the outset that my interpretation and application of this challenge differs from Levins'. Specifically, Levins lamented that the output of brute force models would be "expressed in the form of quotients of sums of products of parameters that would have no meaning for us" (p. 421). This is an issue about how we make sense of the model output given the abstract form that output takes. As already argued in Chapter 2, the output of complex simulation models has become more comprehensible over time as techniques have improved for analysing large data sets. For example, visualisation techniques can be used to convert abstract model output into a representational form that even non-experts can easily understand. However, even if these techniques assist with our understanding of model results, the problem of understanding *how* those results were produced remains.

As I argued in Chapters 2 and 3, brute force models are not typically oriented toward the goal of fostering scientific understanding of key dependency relationships. Models oriented toward that goal are best built with simpler model descriptions and simpler structures. However, the knowledge that comprehension is being achieved elsewhere in the scientific ecosystem does not mean the challenge of comprehensibility facing brute force models can be ignored. Even if brute force models are not used to understand complex targets, they are themselves complex structures that must be understood by modellers in order for modellers to improve their performance and interpret their results. For example, if modellers find their models fail to accurately represent precipitation patterns in the tropics, they must identify which components of the model are responsible for the mismatch such that they can alter those components and improve the model's representational fidelity. Likewise, if a model is used to inform the kinds of interventions that might be made upon the target system, then understanding the model and its limitations can inform the proper application and interpretation of the model. Confidence that the model provides useful information about the target requires understanding both the model and the target.

As stated above, intercomparison projects like those associated with CMIP play an important role in assisting modellers to better understand the causal structures of their simulated systems. In part, this is because these projects provide a set of experiments that detail ways in which models can be varied and which can, therefore, provide information about the causal structure of the model. Consider the experiments used in the MIP for cloud feedback, which direct modellers to decouple their representation of the atmosphere from a realistic representation of the oceans, instead choosing to represent the Earth as an aquaplanet, completely covered with ocean (Webb et al., 2017). This eliminates alternative possibilities and focuses attention on the performance of the remaining components. This investigation of models via systematic variation should be familiar from the discussions of Chapter 3.

Although intercomparison is key to improving model structures and managing errors, it brings with it a comprehensibility problem of its own. Most obviously, if models produce model data in many different formats, this will increase the time and effort researchers need to invest in interpreting data sets, hindering cross-model comparison of model data. If data sets are standardised, then interpreting and analysing different data sets becomes much easier for researchers (Pascoe, Lawrence, Guilyardi, Juckes, & Taylor, 2019). Once again, CMIP has been involved in coordinating such improvements, with Veronika Eyring and colleagues (2016) stating that standardising model data formats is one of CMIP's key tasks. This task is undertaken through the Earth System Grid Federation data replication centres that are that are also tasked with the data management discussed earlier. Within the archival infrastructure established to house model data, a standardised form for this data assists with comprehensibility of model results (Sun et al., 2020).

Standardisation assists in enabling researchers to analyse the results of different models and assists in model intercomparison. But model intercomparison itself is the best route climate scientists have for improving their collective understanding of model behaviour and results (though see Touzé-Peiffer, Barberousse, & Le Treut, 2020 for an alternative view). CMIP is one notable organisation responsible for coordinating these sorts of studies but there are also non-CMIP intercomparison projects. As stated earlier, intercomparison projects are computationally costly, but CMIP has also

responded to these demands by making these studies more accessible for research centres with fewer computational resources at their disposal. While these projects do not solve the problem of comprehensibility, they can at least be used to provide insight into some of the key areas of uncertainty regarding internal structures or provide direction into where uncertainties might be found.

6.3.4 Summary

In this section, I've described CMIP, a community organisation, and the ways in which it has assisted in addressing the epistemic challenges facing brute force models. First, it assists with data hunger by managing a data-archival network which stores both empirical data and model data in standardised forms, allowing for easy access and use by modellers. Second, it coordinates model inter-comparison projects which investigate the representational choices made in the face of computational limits. CMIP has also recently reorganised its approach to these projects to enable modelling centres to contribute to these projects without committing to the computational costs of participating in every project. Finally, archiving model data in a standardised form and running intercomparison projects is one method used by climate scientists to improve their understanding of the internal structures of their models, though it may not be the best method for understanding why models are biased in the ways that they are. It is very good at detecting biases in models, but since it leaves climate models as black boxes, it is not well-suited for addressing what representational choices are responsible for inaccuracies (Touzé-Peiffer et al., 2020).

Are the three epistemic challenges facing brute force modelling completely solved? No. But CMIP and the infrastructure around it go a long way to mitigating the effects of these challenges and serve as an example of what is required to perform brute force modelling as well as possible.

In the next section, I want to look at a contrasting case where many of the features of climate modelling, including organisations like CMIP, are missing. I will examine agent-based models (ABMs), which have been my central focus across the last two chapters. We will see that a lack of standardisation and community infrastructure mean the challenges of brute force modelling go largely

unchecked. However, there are two deeper problems for brute force ABMs. First, the targets they describe differ in so far as data about these systems can be particularly difficult to acquire and the relationship from model to data is often hard to determine. Second, heterogeneity among the systems these models represent divides resources and attention, so the singular international and inter-group focus on models of the one system, like that found for climate models, is absent.

6.4 ABM and the brute force approach

At first, the proposal that ABMs are sometimes instances of the brute force approach might seem surprising. This is because many ABMs are oriented toward theoretical exploration rather than direct representation of worldly target systems, and many agent-based modellers consider theorising to be the proper domain of ABM. Prominent ABM builder Robert Axelrod, for example, states that (1997, pp. 4–5 emphasis added):

> Although agent-based modelling employs simulation, it does not aim to provide an accurate representation of a particular empirical application. *Instead, the goal of agent-based modelling is to enrich our understanding of fundamental processes that may appear in a variety of applications.*

Modellers from other fields continue to hold this view. Social psychologists Eliot Smith and Frederica Conrey, in their paper on using ABM to build theory in their field, claim that (2007, pp. 98–99 emphasis added):

> As we have said, it is crucial to keep in mind that *ABM is a representation of a theory* (typically with fewer than half a dozen fundamental principles), not a representation of messy social reality.

Despite what ABM theoreticians such as these seem to believe, not all ABMs are oriented toward the exploration of theoretical principles. Some are instances of target-directed modelling oriented toward forecasting and intervention. In business, policy, and military, for example, target-directed models are far more common than theoretical ones (Heath et al., 2009). While these models are, I concede, typically not quite as comprehensive as high-fidelity climate models, and so may not be

instances of the brute force approach, they do aim for a moderate degree of realism, so still face the epistemic challenges of brute force modelling even if to a lesser degree. To demonstrate the kinds of ABMs used for this kind of target-directed modelling, let me introduce one final case study in section 6.4.1 before discussing how ABMs face the problems of brute force modelling, especially the problems of data-hunger and comprehensibility.

6.4.1 Modelling Agriculture

One policy area in which ABMs have been used is agriculture policy. Here, I will provide a concrete example of one such ABM, used to represent land use change in New Zealand in response to a price on carbon. My reason for presenting the details of this model is that it is representative of the sorts of ABMs that might contribute to policy decisions. To implement the lessons of the previous chapters, I will do my best to keep the description of the model in line with the ODD protocol described in Chapter 5 and, in particular, the Summary ODD. As you read its description, you will see that some aspects of the model are closely connected to empirical data, such as the information about land use and productivity zones, while other aspects of the model—aspects that are just as important—are based on assumptions lacking the same empirical basis. As I will argue later in the chapter, getting good data and a clear understanding of decision-making processes is a serious constraint for ABM.

New Zealand has had an emissions trading scheme since 2008 but, naturally, policymakers wish to know how emissions can be reduced in line with policy while also minimising any harm to their local industry. In their (2015) paper, Fraser Morgan and Adam Daigneault present an ABM to investigate how New Zealand's climate policy might affect its agricultural and forestry industry, particularly how land is used. As they put it, **The overall purpose of the model** is to "explore the spatial effects on land use, through changes in a [greenhouse gas] price on the forest and agricultural sector, and the effects of information diffusion through farmer social networks" (p. 3)

The authors do not explicitly state **which patterns are used to evaluate the realism of their model**. Nevertheless, we can make some judgments about which patterns could be used given the variables included in the model and presented in the results section. The variables included are farm

net revenue, greenhouse gas emissions and net greenhouse gas emissions, nitrogen leaching, phosphorus loss, and four uses for land: carbon forestry, dairy, forest, and sheep and beef production. So, if any of these appear suspiciously high or low given prior knowledge, then the realism of the model may be called into question. To give an example, one result the authors produce for their test region, the Hurunui and Waiau catchments to the north of Christchurch, is that dairy operations expand from 16,900 ha to 106,472 ha in over a 50-year period. As just suggested, this result could be seen as rather high and bring the realism of the model into question, but the authors argue that the amount of land dedicated towards dairy in the Canterbury region in which the Hurunui and Waiau catchments are located increased "by 172% between 1996 and 2008, and it is projected to expand by an additional 51% by 2020" (p. 10). Consequently, the model results continue, rather than diverge from, historical trends.

The model includes the following *entities*: the landscape, farms, and farmers. The landscape represents the initial cadastral boundaries (property boundaries), the land's current use, and productivity zones, which indicate what enterprises could profitably be pursued on that land. A farm agent is created at the centre of each parcel of land, which sets the properties of the farm as indicated by the landscape layer. The farm agent then creates a farmer agent, which possesses attributes relating to economic and social features along with the decision-making framework that will determine how the model states evolve over time. At initialisation of the model, each farm agent is sorted into an enterprise category, such as pine plantation or dairy. The model then represents how farms change given the price on carbon, what the farm is apt to produce, and the social networks of the farmers.

The most important *processes* of the model, which are repeated at every *time step* are the evaluation of farm profitability by the farmer, which may result in the farmer changing the farm's enterprise. To do this, the model provides each farmer with the expected net revenue for all possible enterprises as well as the option which maximises net revenue. The farmer compares this option to the actual farm enterprise, receiving the money if they are already pursuing this enterprise. If they aren't, then the decision process to change enterprise is initiated. This is a stochastic process, where the chance of change is determined by a baseline value (.2 probability of change), modified according to

features of the farmer's two networks and two further constraints. One network consists of nearby farmer agents (the closest ten) who pursue the same enterprise, while the other consists of geographically adjacent farmer agents. When deciding whether they will change enterprise, farmers consider how well their farm is performing relative to the farms in both their networks, with the costs of change represented by modifications to the probability of change. If their farm is performing better than the mean of the farms in their network on a measure of profitability/ha, then they are less likely to change. If their farm is performing worse, then they are more likely to change (probability 0.1 either way in the case of their like-enterprise network, and 0.05 either way in the case of their neighbourhood). The further two constraints are that farmers who elect to pursue forestry or carbon forestry are locked into this option for 25 virtual years. The second constraint is that, if a farmer agent receives the recommendation that dairy would be more profitable than their current enterprise, there is 75% reduction in the chance that the farmer will change farm enterprise. This is intended to represent the upfront cost to changing from forestry, carbon forestry, or sheep and cattle farming to diary along with the costs of lifestyle changes. According to the modellers, no such barrier exists in the case of changing between forestry, carbon forestry, and sheep and cattle farming. This is based on the modellers' assumption that the costs of changing between these are negligible for the purposes of the model, although a satisfactory explanation for this assumption is not provided.

The final process occurring every time step is that farmers who have reached the end of the farming lifecycle sell their farm if they have not found a successor. Note that the farming lifecycle is based on the work of (Burton, 2009), which includes both the farmer and their successor and five key stages along this cycle for both. The five stages start with accepting leadership of the farm, then move through consolidation, business expansion, the transition of responsibilities toward the successor, and finally retirement. For the successor, the corresponding five stages are birth and socialisation, working full time on the farm, business expansion, the transition of responsibilities toward them, and finally taking leadership of the farm. Each time step of the model represents 5 years of real time, which is intended to reflect the time cost of changing enterprise and roughly the time it takes to produce a yield

from any given enterprise. This also maps onto the five stages of the farming lifecycle well enough for the purposes of the model.[63]

The most important *design concepts* of the model are the basic economic principle that farmers will act to maximise expected utility, best approximated by financial payoff, and interaction between agents. Although famer agents never move, they interact to the extent that they observe one another's profitability/ha and use this information when deciding whether or not to change their enterprise.

In case you are interested, the result of the model, at least for the regions of Canterbury simulated, was that dairy would increase drastically, as mentioned above. This leads to a minor increase in emissions over time—minor because most of the conversions come from high-emitting sheep and cattle farming. Due to the increase in dairy, N leaching and P loss also increase. Given these results, a policy maker may infer that New Zealand's agricultural sector will continue to thrive under a simple price on carbon but that such a policy mechanism will not achieve the goal of reducing emissions, and further policies would be required to address the environmental problems of N leaching and P loss.

The description of this ABM should provide you with a clear idea of the sorts of ABMs that are used to inform policy decisions. Note that I have chosen it because I think it is a representative model, not because I think it is a particularly good model. Note also that my intention is not to provide a detailed critique of the model, although some of my views on this exemplar will find their way into my analysis below. With those caveats in place, let us turn to my discussion of ABMs and brute force modelling.

[63] I say "fairly well" because it is unlikely that the time between stages *birth and socialisation* and *working fulltime* on the farm will be five years. At least, not in contemporary and near-future New Zealand.

6.4.2 ABMs and data-hunger

The first challenge facing brute force models, which should also face target-directed modelling more broadly, even if to a lesser extent, is that of data-hunger. Here we meet a particular problem for ABM, which elsewhere have been recognised to have a tortured relationship with data (Fagiolo, Windrum, & Moneta, 2006; Heckbert, Baynes, & Reeson, 2010; Villamor, Troitzsch, & Van Noordwijk, 2013; Windrum, Fagiolo, & Moneta, 2007). Even if it were true that ABMs were typically or exclusively oriented toward the exploration of theoretical principles, this might lessen the problems of unclear connections between model and world, but it would not remove the problem in its entirety. After all, to the extent that scientific theories should assist in the explanation, prediction, and manipulation of target systems, there should be *some* way of connecting them to the world (Müller et al., 2014, p. 158). Moreover, an unclear connection between model and data is a problem when it comes to model validation. While one aspect of theorising is the exploration of dependencies that may possibly be operating in the target domain, we ultimately wish to isolate dependencies that are probably or actually responsible for observed phenomena. Validating a model with some data will play an important role in elevating a piece of modelling from mere hypothesis to well-confirmed theory. Philippe Giabbanelli and colleagues (2021) conducted a review of 32 ABMs of obesity, which were mostly theoretical, finding that 56.3% of the studies failed to show that the model had been validated.

The primary barrier holding ABMs back from reaching the sort of realism that we see in climate models is a lack of data. Just as in the case of climate models, ABMs aiming toward a high degree of realism will suffer from heavy data demands. The New Zealand land use change model, for example, needs data about how land is currently used and how it could be used. Upon initialisation, contemporary data sets recording actual farm use and productivity zones in the regional target are used to determine the attributes of the farm agents, and records of land use change in the past are used to evaluate whether projected trends are plausible or implausible. The challenges for collecting this data resemble the challenges facing collection of climate data, particularly when climate models include modules dedicated toward representing land use.

With respect to the problem of unclear connections between models and data, the element added with Laatabi et al.'s (2018) extension of the ODD framework, the ODD+2D, introduces four subsections to record an overview of the section, the structure of the data used, a mapping between parts of the data structure and parts of the model, and the patterns observed in the data that should or need not be recreated in the model. If the ODD+2D framework gains widespread adoption as the base ODD framework has, then the problem of an unclear relationship between data and models will be somewhat reduced. That is, it will make transparent what the connection between model and world is *supposed* to be according to the modeller. This is better than nothing. Unfortunately for ABM, however, some deeper difficulties remain on this front that cannot be solved with anything as simple as a best practice of model communication.

There is a deeply difficult aspect of the challenge of data hunger that is particular to ABMs, which is that ABMs like the one described in section 6.4.1 deal explicitly with human decision making. Collecting data about land use, presence of this or that chemical, and even organism abundance is one thing, collecting data about how people think and why they act the way they do is quite another. In the model of New Zealand's agricultural land use change, the farmers' decision-making has little to no connection to empirical data.[64] For example, little justification is given for why network effects modify the farmers' probability of enterprise change as they do and do not take some other values. Indeed, the farmers' decision process determines the model's evolution, so the lack of empirical grounding in this process removes much of the model's predictive power. Basing the transition rules of a model on mere assumptions, as opposed to assumptions with clear empirical grounding or theoretical precedent, is a feature we would expect of a theoretical model intended to

[64] I am mindful that this is an exaggeration. There are basic theoretical principles which are frequently incorporated into models of behaviour and which are also empirically supported. These include, for example, the observation that agents often do what is in their interest, or that agents suffer losses far more than celebrate wins. These principles, however, remain very general. Empirically informed assumptions about detailed aspects of agents' decision-making processes remain far scarcer.

demonstrate the consequences of assumptions or to illustrate simple causal relationships, not one intended to forecast the likely consequences of policy interventions.

Having this data is essential not only to inform the construction of the model, but also to validate the model's assumptions. In the New Zealand land use model, for example, we have no way of validating the model's assumptions about farmers' decision-making procedures. Do farmers really care half as much about how profitable their farm is compared to their neighbours as they do about how profitable it is compared to other farmers pursuing the same enterprise? Do they really have access to precise financial information, like profitability per hectare, about other farmers in their networks? Without any empirical data, we have no way to answer these questions and assess whether the assumptions of the model are any good.

The challenge of data-hunger, then, presents a serious epistemic barrier for brute force (or brute force-like) ABM. It is possible that the increased, perhaps even disturbingly ever-present, abundance of technology with the capacity to monitor and collect data on humans could fill some of the data gaps. While this will undoubtedly go some way to filling gaps in data, we are likely to find that it is much easier to collect data on some aspects of behaviour than on others. It's one thing to collect data on peoples' movements in space and their spending habits, it is another thing to collect data on how they decide to change direction within their careers. Further complications arise when we consider that some of the key decision-making entities, like governments and corporations, are themselves collections of people with their own internal governance structures that can change over time. Again, it's not obvious what data sets should be mined to provide insight into these processes. One way or another, modellers will need to acquire more data about their targets, including human behaviour and decision processes, if they are to build highly complex and realistic models. This may very well be among the biggest challenges facing brute force ABMs.

Before moving on to some of the other barriers between agent-based modelling and brute force modelling, it's worth returning to my original criticism of the decision-making process of the New Zealand land use model, which was that it contains assumptions that are based neither on good

data *nor well-established principles*. Contrast this with climate models. The dynamical core driving the circulation of the atmosphere and ocean components is based on well-established theoretical principles, encoded with the laws of fluid dynamics, which are approximated within the model. In the case of many ABMs, there are no well-established principles or formal descriptions of principles that modellers can use to inform the construction of their models.

6.4.3 ABMs and computational limits

I will say only a little about ABM and the computational limits with which high-fidelity target-directed modelling must grapple. This is simply because, although it is a challenge ABMs will face if modellers attempt to make them as comprehensive as possible, it is not as much of a deep difficulty as the problems of data-hunger and comprehensibility are for ABMs.

Here is one thing that I will say. Melvin Lippe and colleagues (2019) discuss the possibility and current limitations of large-scale—or, more accurately, multi-scale—agent-based modelling of social-ecological systems. Their argument is that modellers should make more complex models that explicitly represent micro parts of the system, such as the people within a government (as opposed to representing the government as a single actor), and macro parts of the systems such as populations adjacent to the focal population which can influence the focal population. As they see it, models that represent a system in isolation and at too coarse a grain can fail to represent them properly because they can be influenced by both factors outside the system and by micro-actors within the system. Such a view is tantamount to advising that ABMs of socio-ecological systems should embrace the brute force approach to improve their representational fidelity. However, this comes with an obvious cost in computational demand. Building multi-scale models as Lippe et al. suggest will require including far more structure in a model. Simply consider representing a government as the individuals that comprise it rather than as a single agent. Moreover, this computational cost would be accompanied by an increased demand for data as these additional structures would have to be calibrated and validated with empirical data.

6.4.4 ABMs and comprehensibility

As discussed in Chapter 5, ABMs are difficult to comprehend. In part, this is because of their complicated structures, which is a general problem faced by computer simulation models and the structures they investigate. More specifically to ABM, however, there is alarmingly little model replication and model communication practices are generally poor. Establishing the sort of standardised model communication frameworks described in Chapter 5 would be a large step forward in improving the comprehensibility of ABMs, which is currently limited by the opacity that poor communication practices create. Beyond implementing standardised communication frameworks and the sharing of source code, all discussed at length in Chapter 5, ABM comprehensibility could be increased through the reuse and recycling of model components. It is this that I will discuss in this section.

Rather than building their models from thoroughly investigated mathematical or causal relationships, or incrementally improving their models over many years, as brute force modellers in climate science do, agent-based modellers have a fondness for building new models from scratch to address the different problems or targets that they wish to investigate (Bithell, 2018). That is, agent-based modellers are neophilic. This tendency to build new models rather than repurposing or improving upon old ones is especially problematic when modellers build the kinds of more complex target-directed models that can be used in policy applications because, in virtue of the greater complexity of their structures, they are more difficult to understand and take longer to investigate. In climate science, by contrast, a relatively small set of structures are investigated for a long time, sometimes multiple decades, and those structures are improved and enriched over time as knowledge and technology advance (Bithell, 2018; Edmonds, 2020). This leads to a better understanding of models because it is easier to investigate a smaller set of model structures, especially if the investigation is permitted to run for a long time (Jeevanjee et al., 2017). Resisting neophilia and more thoroughly investigating existing structures, then, is required for the development of pieces of theory that can be appropriately applied within target-directed models.

However, as with problem of data-hunger, there are features of the targets of ABMs that make this approach, which has been successful in climate modelling, far more difficult. Indeed, these features of ABMs also explain why there are no large intercomparison projects in ABM science. There is a key asymmetry between ABM and global climate modelling: ESMs and GCMs are targeted at one big problem, while ABM is targeted toward many small problems. Of course, this is asymmetry is partly explained by: (1) agent-based models being a far broader category than Earth system models, and (2) agent-based models being distinguished on the basis of the methodology and Earth system models being distinguished on the basis of their target. Nevertheless, the targets that brute force ABMs face are designed to represent, such as land use and energy transitions, are primarily investigated as local and regional issues. Examining the impact of climate policy on New Zealand's agricultural and forestry sector, for example, will be of little interest to those outside of New Zealand. In contrast, a global climate model built by New Zealand's National Institute of Water and Atmosphere would be of interest to Polish researchers qua representation of the global climate. A model of New Zealand's possible economic pathways to lower emissions will likely be of interest to say, the Polish, to the extent that the lessons learned from the model, or the model itself, can be generalised beyond New Zealand, which is possibly not at all.

In Chapter 2, I argued that target heterogeneity was not necessary to create modelling trade-offs, demonstrating that model and target complexity was sufficient to do the job. However, target heterogeneity once again raises its head as something standing in the way of producing brute force models because epistemically respectable brute force modelling requires that resources be concentrated in order to investigate brute force models. Target heterogeneity divides resources, however. In the case of ABM, we do not find a host of different modelling groups researching and representing the same target system and contributing to the same set of policies. Intercomparison projects require a shared target rather than just a shared modelling approach. Climate science might be unique in targeting one system—the Earth's climate—that is of extreme interest to very many researchers, policy makers, and lay-people. While this provides their models with unrivalled

complexity and detail, it also leads to the construction of a set of different models that can be compared.

Is there any solution to improving model comprehensibility if ABM science is characterised by pursuing many smaller problems rather than investigating a few big problems where research resources can be pooled? One possibility that has been raised by agent-based modellers is that modellers should try to avoid building bespoke models, possibly reinventing the wheel, and that modellers should instead attempt to take advantage of the modular nature of ABMs, slowly building up a standardised set of algorithms for representing processes that may appear across a range of targets (Richiardi, 2017). In her short paper, Lynne Hamill (2010) argues that ABM science would benefit from having a set of basic building blocks for the sorts of systems and problems ABMs are intended to address. Since simulation models are built out of sub-components, modellers could have a set of basic building blocks from which they build their models, with some pieces here and there. But, if these models include many already well-understood and investigated computational functions, then those modellers and others should have a better understanding of the whole model structure, at least when compared to modellers looking at an entirely new model structure.

Reusing and recycling model structures and their components is not without its own problems, however. In their paper on ABM and policy making, Bruce Edmonds and Lia ní Aodha (2018) describe two problems facing ABM in the policy context. One problem is that a model can become accepted and spread to new applications where it may be less apt. Another is confusion over model purpose, and they argue that models should be published with a clearly stated purpose since this is essential to an evaluation and correct application of that model. Although these are real problems, they can be mitigated through the use of a framework like the ODD and its extensions, which requires modellers to reflect on their modelling choices and explicitly state the purpose of the model and patterns used to evaluate the model. One optimistic possibility is that engaging in the process of constructing ODDs will force modellers to reflect on their unhelpful application of a model structure and self-correct. Alternatively, if motivations and purpose are provided in a transparent format like the ODD, critics will be able to isolate the problematic assumptions in the model relative to its new

context more quickly. Consequently, reusing model structure to improve the epistemic state of ABM science will succeed only in concert with improved practices of communicating model structures and modelling choices.

6.4.5 Summary

In this section, I have argued that the challenges facing brute force models appear far worse for ABMs than they do in the climate case. This is because, as argued in Chapter 4 and demonstrated in the example of section 6.4.1, ABMs focus on agent rules and behaviour. Data about this behaviour and, in particular, these decision rules, is particularly difficult to gather. ABMs also face comprehensibility challenges related to a tendency within the ABM community to build new models for each new application, reducing the capacity for modellers to grow their knowledge about a model over time and over multiple iterations. Likewise, there are few instances where many modellers and research groups are focused on a shared target, once again reducing the capacity for modellers to pool resources as they can in the climate case. Although I argued, in Chapter 2, that target and model complexity was enough to generate modelling trade-offs and that heterogeneous targets were not required, this section has demonstrated that target heterogeneity places serious roadblocks in the way of progress in the context of brute force modelling.

6.5 Conclusion

In this chapter, I have argued that the three epistemic challenges facing brute force models can be addressed, and the epistemic situation brute force modellers find themselves in improved, through community infrastructure that coordinates the extensive investigation of brute force models. The solutions to the problems of brute force modelling are to build institutions and infrastructure directed toward managing the three epistemic challenges. First, institutions are required to collect and store data in a fashion that ensures it is widely available to those in the scientific community who need to inform and assess their modelling efforts. Second, institutions are required to assess the impact of

approximations and idealisations introduced to models in the face of computational constraints. And, third, model descriptions and explanations of modelling assumptions and choices must be detailed, standardised, and made freely available to allow for improved model interpretation by other modellers and members of the scientific community. While the value of something like CMIP might seem obvious now, before the creation of the project, climate modelling was carried out independently by different modelling groups. In part, such organisation was a natural consequence of data sharing limitations faced by scientists at the time. Indeed, even when the CMIP took its initial steps, limitations of computer power and networks led to a relatively modest goal for CMIP. From that perspective, the three epistemic challenges Levins described might have looked insurmountable at the time he was writing.

Despite the success of brute force modelling in climate science, this chapter's look at target-directed ABMs demonstrates that not all fields of study lend themselves to the extensive concentration of resources that CMIP and its surrounding efforts comprise. Mitigating the challenges of brute force modelling comes at great cost. It takes many minds concentrating on problem solving, many eyes fixating on model data and studies, many computers running the simulations and storing the model data, and all sorts of tech to collect the data in the first place. Indeed, state-of-the-art climate modelling is, perhaps not unique, but uncommon, in so far as it is a situation where very many researchers are attending to the one system and are provided with lots of resources to do this because of the payoff to governments of getting information about the likely and possibly behaviours of this system. In other situations, fewer researchers are interested in a shared target system, so their resources are divided.

To the extent that we lack brute force ABMs, this lack is not the result of a direct choice, but of the constraints modellers face, some of which they share with climate modelling and some of which are unique to ABMs. Most strikingly, brute force ABMs would have an appetite for data about human behaviour and decisions that that is particularly difficult to acquire, even with technological advances. With respect to the comprehensibility of ABMs, this would greatly be improved by implementing the kind of standardised documentation and code-sharing frameworks described in the previous chapter. While the systematic intercomparison of models is unlikely to be as successful in ABM science as it is

in climate science due to the heterogeneity among target systems, some are optimistic that a set of basic modules could be devised which would be generally applicable and from which more complicated models could be constructed. These basic modules would be well understood and would form a commonality among otherwise disconnected models, but at present this is optimistic speculation.

A general lesson for the epistemology of computer simulation that can be derived from this chapter relates back to Levins' sceptical perspective on brute force modelling. Recall from Chapter 2 that Levins was sceptical about the brute force strategy because of the multiple functions that models can play, some of which brute force models are poorly suited towards. In particular, brute force models are unable to achieve generality due to their specificity and the heterogeneity among target systems. Consequently, general models must be far less detailed than brute force models. This chapter has demonstrated that heterogeneity among target systems presents a problem for the brute force approach not discussed by Levins in his work. This is that the resources required to pursue the brute force approach in earnest may only be available in cases where heterogeneity is absent because heterogeneity splits the efforts of brute force modellers, reducing their ability to undertake the thorough investigations of their models and to provide the infrastructure required to address problems like the challenge of data-hunger. Heterogeneity then, not only generates modelling trade-offs, as Levins described, but divides resources, suggesting that, in cases where target heterogeneity characterises a scientific field, brute force modelling may not be among those strategies that can be successfully pursued.

7

Conclusion: Lessons for the epistemology of computer simulation

At the outset of this thesis, I stated that I would endeavour to contribute to the philosophy of science and, specifically, the philosophy of modelling and computer simulation by investigating the epistemology of computer simulations. To do this, I focused on two fields of study: climate science, which has long made use of state-of-the-art computer simulations, and agent-based models (ABMs), which are very different models aimed at representing populations of interacting agents but which have a history of underperforming. In this final chapter, I wish to draw the lessons of the thesis together to make general statements about the epistemology of modelling and computer simulation.

Before I begin presenting my lessons, however, let me return to the basic framework of scientific models that I presented in Chapter 1. This basic framework captures both simulation and non-simulation models, with the primary features of computer simulation models being that their structures are investigated through the use of digital computers that manipulate the structures' formal description, which is an algorithm written in some programming language. Although this is a very broad definition of both model and computer simulation model, as I will demonstrate, it is still useful for analysing the epistemology of computer simulation.

First, if you build a model with a mathematical structure and do *not* use a digital computer to investigate that structure, then you are rather limited with respect to how complicated and complex you can make your structure. If you specify your model with sets of continuous mathematical equations and plan on analysing those equations with the formal machinery such mathematics provides, then you cannot investigate many complex systems because there are no known analytic solutions to many of these equations. If you specify your model with discrete mathematics, then you are limited by the amount of time available to carry out the calculations. Consequently, using digital computers to perform your calculations permits the construction of models with structures that are

vastly more complicated and complex than the structures of those models investigated without computers. Hence, the first hallmarks of computer simulation models are the sizes of their model structures and counterpart model descriptions.

Given that computer simulation lifts the ceiling on the size of mathematical model structures that can feasibly be investigated, it may be tempting to take full advantage of this capacity and to make models with very complex structures indeed. The arguments of this thesis, particularly Chapters 2 and 6, suggest that this temptation should, in many cases, be resisted. This is because increasing the size of a model structure creates epistemic challenges for which one must be prepared, such as having the empirical data to ensure that the model structure remains a faithful representation of its target relative to the model's purpose, having the human resources required to evaluate the representational choices made when building the model, or having the computational resources required to run the simulations that are used in that evaluation process.

In addition to these practical problems, models with big structures are also poorly suited to the task of fostering scientific understanding. This is both because they are difficult to understand themselves, so scientists struggle to get a grip on them, but also because they do a poor job of isolating key causal dependencies. Rather, they are well suited to investigating some specific target or narrow set of targets for the purpose of forecasting or intervention. Nevertheless, these tasks are performed best when scientists have already gained enough understanding of these targets that they can evaluate their complex models and ensure that their projections and predictions are reliable given their purpose.

The philosophical framework of scientific models I adopted in Chapter 1 also includes a distinction between the formal descriptions of models, such as sets of equations and lines of code, and the mathematical relationships that these formal descriptions specify. I used this distinction to define computer simulation models as models investigated through computational manipulation of the model description. And, while some are sceptical of such a distinction, the arguments of this thesis demonstrate that this distinction enables an easy discussion of a serious source of uncertainty for computer simulation models. As suggested above, as one increases the size of the model structure, this

structure becomes more difficult to comprehend as does its formal description. To an expert mathematician, the connection between a set of equations and a set of mathematical relationships is transparent. To the expert simulation modeller, however, there is no such transparency. Modellers must be careful then, to ensure that the formal description being manipulated in their simulation studies *actually* specifies the mathematical relationships desired rather than some other set of relationships.

If you increase the size of your model structure, it also becomes more difficult to ensure that the formal description specifies the relationships desired. Likewise, it becomes more difficult to determine whether the relationships included in the structure behave as desired and there is not some other set of relationships that would have been better to include given the purposes of your model. Reducing this uncertainty and ensuring that the big structure specified was a good one to specify in the first place is a costly enterprise on its own. As the arguments of Chapter 6 demonstrate, it takes many expert eyes with a lot of time and resources to reduce this uncertainty. Even then, the uncertainty is not eliminated completely. Given this, modellers should consider whether they have the time and resources required to address this uncertainty as best they can or whether they should opt for simplicity wherever possible. This is especially important when the mathematical relationships are intended to encode particular assumptions or hypotheses of the modeller. If the modeller unknowingly specifies some other relationships, then they may not investigate their assumptions and hypotheses at all, but some nearby assumptions and hypotheses.

The uncertainty between specifications and structures also has an impact on the ability for modellers to investigate and understand the structures built by their colleagues. A prescriptive argument I made in Chapter 5 is that modellers must provide their source code for simulation models because the kinds of brief natural language descriptions that typically appear in manuscripts or published papers simply do not contain enough information for modellers to identify the relationships being investigated and to specify them for themselves. However, it would be even better if the source code was accompanied by *detailed* natural language descriptions of each model element and an explanation of the motivation behind the modelling choice made. This is useful for two reasons. First,

a modeller who may be unfamiliar with a particular source code or programming language can make use of the detailed natural language description to interpret it. Second, even if a modeller is familiar with the source code language, a detailed natural language description of the structure and motivations behind representational choices can be used to assess whether the structure specified with the source code achieves the aims set out by the modeller in the detailed natural language description.

If the first epistemic positive that comes with computer simulation is that larger and more complex mathematical structures can now be investigated—a positive that comes with a note of caution, however—then the second epistemic positive is that, within these more complex structures, are different *kinds* of structures that have different representational capacities. That is, there are structures that can only be feasibly investigated with computer simulation that are especially well suited to representing certain kinds of targets.

Chapter 4 described in detail how ABMs are especially well-suited to representing populations of interacting agents. A great many areas of science, and many areas of policymaking, are concerned with the collective behaviour of populations. Also, the philosophy of science for at least the last 20 years has emphasised the value of representing complex systems at the level of the individual components comprising the system. ABMs are valuable because they allow for the representation of populations—itself nothing new—but by representing the individual members of the population rather than representing properties of the system as a whole. This allows for the construction of models that can exemplify dependency relationships between the behavioural rules governing the members, the features of a heterogeneous environment, and the complex and often surprising behaviour of the system as a whole.

As before, access to these new mathematical structures comes with a note of caution. Without careful examination, it can be difficult to determine exactly what assumptions are responsible for the resulting behaviour of the model. Particularly concerning for their ability to explain, Chapter 6 demonstrated that the relationships between ABMs and empirical data about their target systems are often unclear. This is partly because ABMs are a relatively recent form of modelling that has not yet

developed proper standards, but also because the kinds of systems they represent are ones for which it is particularly hard to get data that would validate or invalidate the model. ABMs function by representing individual members of a population and the rules that govern their behaviours, but it is very difficult to know what decision rules actual populations are employing even to a first approximation. This is a serious epistemic limitation for ABMs with no clear solution other than limiting their use to contexts where a close fit between model and target is less important, such as if the models are only used to explore possibilities or to assist with reasoning through theoretical assumptions, or to pursue a closer fit between model and target only in those contexts where decision rules are well known at least to a first approximation, such as contexts where, for example, consumers are assumed to want to save money. Alternatively, it may reveal that ABMs are valuable when it is the environmental structures, rather than the agent rules, that carry the largest portion of the explanatory weight, since the representation of environmental structures *is* straightforward to evaluate.

Although the epistemic value of modelling is improved by using multiple models with different assumptions, there is decreasing value if the number of models increases too quickly and the relationships between models are not tracked. A team of models is useful, a crowd or riot of models is not. While model proliferation may seem like progress because the space of possible models is explored more comprehensively, as Chapter 5 argued, the quality of the exploration can sink to the point where little of value is gained from any one model. Rather, the anarchic nature of the exploration limits any value that might have otherwise been gained from the models as a collection. Since the epistemic value of models is improved as more scientists can perform studies of their own on a model structure of family of related model structures, and since the capacity for this to take place decreases the number of independent models increases, model proliferation impedes the careful study of existing models.

Highly complex models, which typically require simulation to investigate, do not generalise well. This is because the details that highly complex models include make them dissimilar to a range of target systems, lacking these features or including incommensurate ones. As Chapter 2 discussed, this creates a trade-off. However, as Chapter 6 argued, heterogeneity among target systems does not

just contribute to the poor generalisability of highly detailed models. Additionally, target heterogeneity can weaken the epistemic status of highly complex models because heterogeneity divides the scientific resources, including resources for gathering and managing data, performing model evaluation and replication, running simulations, and so on. When a large number of modellers share a target or small set of similar targets, they are able to pool these resources and perform the epistemic practices that are required to increase the reliability of complex models. Unless strategies are developed to find the commonality among the heterogeneity, the epistemic value of complex models in fields characterised by the study of heterogeneous system may remain severely constrained.

I wish to end with a mention of four topics where I think questions remain and on which my future work might focus. The first topic that deserves more attention is that of the connection between scientific understanding and robustness analysis. As discussed in Chapter 3, there is an extensive literature on robustness analysis in the philosophy of science, and there is a growing literature on scientific understanding. However, little work has been done explicitly connecting robustness analysis and understanding, and I have argued that we have good reasons to think that this practice and this epistemic achievement are closely related. A more thorough examination of the practice of robustness analysis and its connection to the insights of both factive and non-factive approaches to scientific understanding is warranted and could improve philosophical accounts of both robustness analysis and scientific understanding.

The second topic deserving more attention is that of the connection between population thinking and explanation. As Chapter 4 argued, population thinking has been contrasted with mechanistic explanation, but, as I demonstrated there, population thinking has no associated explanatory framework. Many kinds of models can be used by those engaging in population thinking. Therefore, more work could be done on the concept of population thinking in the context of explanatory frameworks. Relatedly, there is more work to be done on the explanatory capacities of ABMs.

Third, the practice of building the sorts of model spaces described in Chapter 3 deserves further examination. It would be valuable to find an area—perhaps an area of agent-based modelling—and to construct some model spaces to see just how powerful they are for systematising the exploration of dependencies and representational choices. Although I believe the arguments of Chapter 3 strongly suggest that such a practice will be valuable, only the application of the practice will definitively demonstrate the extent of its utility.

Finally, this thesis ended on a somewhat pessimistic note: fields characterised by target heterogeneity often lack the resources required to face the epistemic challenges associated with high-fidelity modelling. The next logical step, then, is a thorough consideration of the methods that can be used to improve the epistemic situation in these cases. For example, perhaps statistical models, which I have not discussed at all throughout my thesis provide powerful insights into the behaviour of such systems while avoiding the epistemic challenges faced by dynamical models. Perhaps they face a set of epistemic challenges all of their own. But I cannot answer these questions here.

Thank you for your time.

Bibliography

Alexander, K., & Easterbrook, S. M. (2015). The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geoscientific Model Development*, *8*(4), 1221–1232.

Alexeev, V. A. (2003). Sensitivity to CO2 doubling of an atmospheric GCM coupled to an oceanic mixed layer: a linear analysis. *Climate Dynamics*, *20*(7–8), 775–787.

Alexeev, V. A., Langen, P. L., & Bates, J. R. (2005). Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Climate Dynamics*, *24*(7–8), 655–666.

An, L., Grimm, V., & Turner II, B. L. (2020). Meeting grand challenges in agent-based models. *Journal of Artificial Societies and Social Simulation*, *23*(1).

Andersen, H. K. (2011). Mechanisms, laws, and regularities. *Philosophy of Science*, *78*(2), 325–331.

Andersen, H. K. (2012). The case for regularity in mechanistic causal explanation. *Synthese*, *189*(3), 415–432.

Anderson, M. L. (2014). *After phrenology* (Vol. 547). MIT Press Cambridge, MA.

Angus, S. D., & Hassani-Mahmooei, B. (2015). " Anarchy" Reigns: A Quantitative Analysis of Agent-Based Modelling Publication Practices in JASSS, 2001-2012. *Journal of Artificial Societies and Social Simulation*, *18*(4), 16.

Attanasio, A., Pasini, A., & Triacca, U. (2012). A contribution to attribution of recent global warming by out-of-sample Granger causality analysis. *Atmospheric Science Letters*, *13*(1), 67–72.

Attanasio, A., Pasini, A., & Triacca, U. (2013). Granger causality analyses for climatic attribution. *Atmospheric and Climate Sciences*, *3*(04), 515.

Axelrod, R. (1984). *The evolution of cooperation*. Basic books.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration* (Vol. 3). Princeton University Press.

Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., … Elkington, M. (2018). Requirements for a global data infrastructure in support of CMIP6. *Geoscientific Model Development*, *11*(9), 3659–3680.

Ball, P. (2008). Cellular memory hints at the origins of intelligence. *Nature News*, *451*(7177), 385–385. https://doi.org/10.1038/451385a

Bankes, S. C. (2002). Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7199–7200.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, *78*(4), 533–557.

Bechtel, W. (2012). Understanding endogenously active mechanisms: A scientific and philosophical challenge. *European Journal for Philosophy of Science*, *2*(2), 233–248.

Bechtel, W. (2016). Mechanists Must be Holists Too! Perspectives from Circadian Biology. *Journal of the History of Biology*, *49*(4), 1–27. https://doi.org/10.1007/s10739-016-9439-6

Bechtel, W. (2017). Systems Biology: Negotiating Between Holism and Reductionism. In *Philosophy of Systems Biology* (pp. 25–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-47000-9_2

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

Bechtel, W., & Abrahamsen, A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, *18*(4), 707–723.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.

Beekman, M., & Latty, T. (2015). Brainless But Multi-Headed: Decision-Making by the Acellular Slime Mould Physarum Polycephalum. *Journal of Molecular Biology*, *427*(23), 3734–3743. https://doi.org/10.1016/j.jmb.2015.07.007

Bennet, F. V. (1970). Apollo lunar descent and ascent trajectories. AIAA 8th Aerospace Sciences Meeting 19-21 January. New York.

Berger, U., Rivera-Monroy, V. H., Doyle, T. W., Dahdouh-Guebas, F., Duke, N. C., Fontalvo-Herazo, M. L., … Piou, C. (2008). Advances and limitations of individual-based models to analyze and predict dynamics of mangrove forests: A review. *Aquatic Botany*, *89*(2), 260–274.

Bithell, M. (2018). Continuous model development: a plea for persistent virtual worlds. *Review of Artificial Societies and Social Simulation*.

Bjerknes, V. (1904). Das Problem der Wettervorhers-age, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteor. Z.*, *21*, 1–7.

Boisseau, R. P., Vogel, D., & Dussutour, A. (2016). Habituation in non-neural organisms: evidence from slime moulds. *Proc. R. Soc. B*, *283*(1829), 20160446.

Bony, S., Stevens, B., Held, I. H., Mitchell, J. F., Dufresne, J.-L., Emanuel, K. A., … Senior, C. (2013). Carbon dioxide and climate: perspectives on a scientific assessment. In *Climate Science for Serving Society* (pp. 391–413). Springer.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Elsevier.

Bromberger, S. (1966). Questions. *The Journal of Philosophy*, *63*(20), 597–606.

Burton, R. J. F. (2009). Strategic decision-making in agriculture: an international perspective of key social and structural influences. *Lincoln, New Zealand AgResearch*.

Calcott, B. (2011). Wimsatt and the robustness family: Review of Wimsatt's Re-engineering Philosophy for Limited Beings. *Biology & Philosophy*, *26*(2), 281–293.

https://doi.org/10.1007/s10539-010-9202-x

Carrier, M., & Lenhard, J. (2019). Climate models: How to assess their reliability. *International Studies in the Philosophy of Science*, *32*(2), 81–100.

Cartwright, N. (1983). How the Laws of Physics Lie. 1983. *Cambridge UP*.

Charney, J. G. (1963). Numerical experiments in atmospheric hydrodynamics. In *Experimental Arithmetic, High Speed Computing and Mathematics. Proc. Symp. Appl. Math* (Vol. 15, pp. 289–310).

Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.

Chisholm, A. (1972). Philosophers of earth: conservations with ecologists. In *Philosophers of earth: conservations with ecologists*. EP Dutton.

Clement, A., Bellomo, K., Murphy, L. N., Cane, M. A., Mauritsen, T., Rädel, G., & Stevens, B. (2015). The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science*, *350*(6258), 320–324.

Craver, C. F. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford University Press.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: discoveries across the life sciences*. University of Chicago Press.

Crooks, A., Castle, C., & Batty, M. (2008). Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems*, *32*(6), 417–430.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy*, *72*, 741–764.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge: MA: Bradform/MIT Press.

Cummins, R. (2000). How does it work?" versus" what are the laws?": Two conceptions of psychological explanation. *Explanation and Cognition*, 117–144.

Currie, A. (2017). From Models-as-Fictions to Models-as-Tools. *Ergo, an Open Access Journal of Philosophy*, *4*.

Dalmedico, A. D. (2001). History and epistemology of models: Meteorology (1946–1963) as a case study. *Archive for History of Exact Sciences*, *55*(5), 395–422.

Day, C. C., Zollner, P. A., Gilbert, J. H., & McCann, N. P. (2020). Individual-based modeling highlights the importance of mortality and landscape structure in measures of functional connectivity. *Landscape Ecology*, *35*(10), 2191–2208.

De Regt, H. W. (2009). The epistemic value of understanding. *Philosophy of Science*, *76*(5), 585–597.

De Regt, H. W. (2014). Visualization as a tool for understanding. *Perspectives on Science*, *22*(3), 377–396.

Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies in History and Philosophy of Science Part A*, *56*, 72–83.

Dellsén, F. (2018). Beyond Explanation: Understanding as Dependency Modelling. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy058

Dowling, D. (1999). Experimenting on theories. *Science in Context*, *12*(2), 261–273.

Downes, S. M. (1992). The Importance of Models in Theorizing: A Deflationary Semantic View. In *Proceedings of the Biennial Meetings of the Philosophy of Science Association* (Vol. 1, pp. 142–153). https://doi.org/10.1086/psaprocbienmeetp.1992.1.192750

Downes, S. M. (2011). Scientific Models. *Philosophy Compass*, *6*(11), 757–764. https://doi.org/10.1111/j.1747-9991.2011.00441.x

Edmonds, B. (2020). Good Modelling Takes a Lot of Time and Many Eyes. *Review of Artificial*.

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., … Squazzoni, F. (2019). Different Modelling Purposes. *Journal of Artificial Societies and Social Simulation*,

22(3).

Edmonds, B., & ní Aodha, L. (2018). Using Agent-Based Modelling to Inform Policy–What Could Possibly Go Wrong? In *International Workshop on Multi-Agent Systems and Agent-Based Simulation* (pp. 1–16). Springer.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*.

Elgin, C. Z. (2017). *True enough*. MIT Press.

Elliott-Graves, A. (2018). Generality and Causal Interdependence in Ecology. *Philosophy of Science*, *85*(5), 1102–1114.

Elliott-Graves, A. (2020a). The Value of Imprecise Prediction. *Philosophy, Theory, and Practice in Biology*.

Elliott-Graves, A. (2020b). What is a target system? *Biology & Philosophy*, *35*(2), 1–22.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.

Eubank, S., Eckstrand, I., Lewis, B., Venkatramanan, S., Marathe, M., & Barrett, C. L. (2020). Commentary on Ferguson, et al.,"Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand." *Bulletin of Mathematical Biology*, *82*(4), 1–7.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958.

Fagiolo, G., Windrum, P., & Moneta, A. (2006). *Empirical validation of agent-based models: A critical survey*. LEM Working Paper Series.

Fairley, R. E. (1976). Dynamic testing of simulation software. In *Proc. 1976 Summer Computer Simulation Conf* (pp. 40–46).

Fang, W. (2017). Holistic modeling: an objection to Weisberg's weighted feature-matching account. *Synthese*, 1743–1764.

Forber, P. (2010). Confirmation and explaining how possible. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *41*(1), 32–40.

Frigg, R. (2010). Models and fiction. *Synthese*, *172*(2), 251–268. https://doi.org/10.1007/s11229-009-9505-0

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese*, *169*(3), 593–613. https://doi.org/10.1007/s11229-008-9438-z

Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., Del Olmo, R., López-Paredes, A., & Edmonds, B. (2009). Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, *12*(1), 1.

Galison, P. (1996). Computer simulations and the trading zone. In P. L. Galison & D. J. Stump (Eds.), *The Disunity of Science: Boundaries, Contexts, and Power* (pp. 118–157). Redwood City: Stanford University Press.

Gao, C., Liu, C., Schenz, D., Li, X., Zhang, Z., Jusup, M., … Nakagaki, T. (2018). Does being multi-headed make you better at solving problems? A survey of Physarum-based models and computations. *Physics of Life Reviews*.

Giabbanelli, P. J., Tison, B., & Keith, J. (2021). The application of modelling and simulation to public health: Assessing the quality of Agent-Based Models for obesity. *Simulation Modelling Practice and Theory*, 102268.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.

Glennan, S. (2017). The new mechanical philosophy. Oxford University Press.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, *21*(5), 725–740.

Godfrey-Smith, P. (2009a). *Darwinian populations and natural selection*. Oxford University Press.

Godfrey-Smith, P. (2009b). Models and fictions in science. *Philosophical Studies*, *143*(1), 101–116. https://doi.org/10.1007/s11098-008-9313-2

Goodman, N. (1976). Languages ofArt: An Approach to a Theory of Symbols (Indianapolis: Hackett).

Gramelsberger, G., Lenhard, J., & Parker, W. S. (2020). Philosophical perspectives on Earth system modeling: Truth, adequacy, and understanding. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001720.

Griffin, A. F., & Stanish, C. (2007). An agent-based model of prehistoric settlement patterns and political consolidation in the Lake Titicaca Basin of Peru and Bolivia. *Structure and Dynamics*, *2*(2).

Grimm, S. R. (2011). Understanding. In *The Routledge companion to epistemology* (pp. 110–120). Routledge.

Grimm, S. R. (2014). Understanding as knowledge of causes. In *Virtue epistemology naturalized* (pp. 329–345). Springer.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., … Huse, G. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1–2), 115–126.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: a review and first update. *Ecological Modelling*, *221*(23), 2760–2768.

Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmański, J., & Wissel, C. (1996). Pattern-oriented modelling in population ecology. *Science of the Total Environment*, *183*(1–2), 151–166.

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., … Groeneveld, J. (2020). The odd protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, *23*(2), 1–7.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hamill, L. (2010). Agent-based modelling: The next 15 years. *Journal of Artificial Societies and Social Simulation*, *13*(4), 7.

Hartmann, S. (1996). The world as a process. In *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77–100). Springer.

Heath, B., Hill, R., & Ciarallo, F. (2009). A survey of agent-based modeling practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation*, *12*(4), 9.

Heckbert, S., Baynes, T., & Reeson, A. (2010). Agent-based modeling in ecological economics. *Annals of the New York Academy of Sciences*, *1185*(1), 39–53.

Hedrich, R., & Neher, E. (2018). Venus flytrap: how an excitable, carnivorous plant works. *Trends in Plant Science*, *23*(3), 220–234.

Held, I. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*(11), 1609–1614.

Held, I. (2014). Simplicity amid complexity. *Science*, *343*(6176), 1206–1207.

Hempel, C. G. (1965). 1965:" Aspects of Scientific Explanation". New York: Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Heyes, C. (2018a). *Cognitive gadgets: the cultural evolution of thinking*. Harvard University Press.

Heyes, C. (2018b). *Cognitive Gadgets*. The Belknap Press of Havard University Press.

Heymann, M., & Achermann, D. (2018). From climatology to climate science in the twentieth century. In *The Palgrave handbook of climate history* (pp. 605–632). Springer.

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., … Bolt, R. M. (2018). Historical (1750-2014) anthropogenic emissions of reactive gases and aerosols from the Community Emission Data System (CEDS). *Geoscientific Model Development*, *11*, 369–408.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, *64*, S325–S336.

Hughes, R. I. G. (1999). The Ising model, computer simulation, and universal physics. *IDEAS IN CONTEXT*, *52*, 97–145.

Humphreys, P. (2004). *Extending ourselves: computational science, empiricism, and scientific method*. Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626.

Jeevanjee, N., Hassanzadeh, P., Hill, S., & Sheshadri, A. (2017). A perspective on climate model hierarchies. *Journal of Advances in Modeling Earth Systems*, *9*(4), 1760–1771.

Jones, J. (2015). *From pattern formation to material computation : multi-agent modelling of physarum polycephalum* (Vol. 15). Springer. Retrieved from http://dx.doi.org/10.1007/978-3-319-16823-4

Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., … Sénési, S. (2020). The CMIP6 Data Request (DREQ, version 01.00. 31). *Geoscientific Model Development*, *13*(1), 201–224.

Kaiser, D. (2009). *Drawing theories apart: The dispersion of Feynman diagrams in postwar physics*. University of Chicago Press.

Katzav, J., & Parker, W. S. (2015). The future of climate modeling. *Climatic Change*, *132*(4), 475–

487.

Khalifa, K. (2020). Understanding, Truth, and Epistemic Goals. *Philosophy of Science*, *87*(5), 944–956.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Hess-Holden, C. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*(5), e1002456.

Knuuttila, T., & Loettgers, A. (2011). Causal isolation robustness analysis: the combinatorial strategy of circadian clock research. *Biology & Philosophy*, *26*(5), 773–791.

Kreuzwieser, J., Scheerer, U., Kruse, J., Burzlaff, T., Honsel, A., Alfarraj, S., … Kreuzer, I. (2014). The Venus flytrap attracts insects by the release of volatile organic compounds. *Journal of Experimental Botany*, *65*(2), 755–766.

Kuorikoski, J. (2009). Two concepts of mechanism: Componential causal system and abstract form of interaction. *International Studies in the Philosophy of Science*, *23*(2), 143–160.

Laatabi, A., Marilleau, N., Nguyen-Huu, T., Hbid, H., & Babram, M. A. (2018). ODD+ 2D: an ODD based protocol for mapping data to empirical ABMs. *Journal of Artificial Societies and Social Simulation*, *21*(2).

Lehtinen, A. (2016). Allocating confirmation with derivational robustness. *Philosophical Studies*, *173*(9), 2487–2509.

Lehtinen, A. (2018). Derivational robustness and indirect confirmation. *Erkenntnis*, *83*(3), 539–576.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 253–262.

Levins, R. (1966). The strategy of model building in population biology. *American Naturalist*, *54*(4), 421–431. https://doi.org/10.2307/27836590

Levins, R. (1968a). Ecological engineering: theory and technology. Stony Brook Foundation, Inc.

Levins, R. (1968b). *Evolution in changing environments: some theoretical explorations*. Princeton University Press.

Levins, R. (1973). The limits of complexity. *Hierarchy Theory-The Challenge of Complex Systems.(Ed: Pattee, HH) George Braziller, New York*, 109–127.

Levins, R. (1993). A response to Orzack and Sober: formal analysis and the fluidity of science. *The Quarterly Review of Biology*, *68*(4), 547–555.

Levy, A. (2013). Three kinds of new mechanism. *Biology & Philosophy*, *28*(1), 99–114.

Levy, A. (2014). Machine-Likeness and Explanation by Decomposition. *Philosopher's Imprint*, *14*(6).

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science*, *80*(2), 241–261.

Levy, A., & Bechtel, W. (2016). Towards Mechanism 2.0: Expanding the Scope of Mechanistic Explanation.

Lewontin, R. C. (1985). Adaptation. In R. Levins & R. C. Lewontin (Eds.), *The Dialectical Biologist* (pp. 65–84). Cambridge: MA: Harvard University Press.

Lippe, M., Bithell, M., Gotts, N., Natalini, D., Barbrook-Johnson, P., Giupponi, C., … Matthews, R. B. (2019). Using agent-based modelling to simulate social-ecological systems across scales. *GeoInformatica*, *23*(2), 269–298.

Lisciandra, C. (2017). Robustness analysis and tractability in modeling. *European Journal for Philosophy of Science*, *7*(1), 79–95.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science*, *77*(5), 971–984.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies*

in History and Philosophy of Science Part A, *49*, 58–68.

Lorscheid, I., Berger, U., Grimm, V., & Meyer, M. (2019). From cases to general principles: A call for theory development through agent-based modeling. *Ecological Modelling*.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

MacPherson, B., & Gras, R. (2016). Individual-based ecological models: Adjunctive tools or experimental systems? *Ecological Modelling*, *323*, 106–114.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*(1), 143–166.

Maher, P., Gerber, E. P., Medeiros, B., Merlis, T. M., Sherwood, S., Sheshadri, A., … Zurita-Gotor, P. (2019). Model hierarchies for understanding atmospheric circulation. *Reviews of Geophysics*.

Manson, S., An, L., Clarke, K. C., Heppenstall, A., Koch, J., Krzyzanowski, B., … Shook, E. (2020). Methodological issues of spatial agent-based models. *JASSS-THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION*, *23*(1).

Matthewson, J. (2011). Trade-offs in model-building: A more target-oriented approach. *Studies in History and Philosophy of Science Part A*, *42*(2), 324–333.

Matthewson, J. (2017). Models of mechanisms. *Routledge Handbook of Mechanisms and Mechanical Philosophy*, 225–237.

Matthewson, J. (2020). Detail and generality in mechanistic explanation. *Studies in History and Philosophy of Science Part A*, *80*, 28–36.

Matthewson, J., & Calcott, B. (2011). Mechanistic models of population-level phenomena. *Biology and Philosophy*, *26*(5), 737–756. https://doi.org/10.1007/s10539-011-9277-z

Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*,

170(1), 169–190.

Mazzocchi, F., & Pasini, A. (2017). Climate model pluralism beyond dynamical ensembles. *Wiley Interdisciplinary Reviews: Climate Change*, *8*(6).

McGuffie, K., & Henderson-Sellers, A. (2013). *A Climate Modelling Primer*. John Wiley & Sons.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (1997). Intercomparison makes for a better climate model. *Eos, Transactions American Geophysical Union*, *78*(41), 445–451.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The coupled model intercomparison project (CMIP). *Bulletin of the American Meteorological Society*, *81*(2), 313–318.

Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: a review of the literature on smart urban governance. *International Review of Administrative Sciences*, *82*(2), 392–408.

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., … Trudinger, C. M. (2017). Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geoscientific Model Development*, *10*, 2057–2116.

Mitchell, S. D. (2009). *Unsimple truths: Science, complexity, and policy*. University of Chicago Press.

Moore, P. B. (2012). How should we think about the ribosome? *Annual Review of Biophysics*, *41*, 1–19.

Morgan, F. J., & Daigneault, A. J. (2015). Estimating impacts of climate change policy on land use: An agent-based modelling approach. *PLoS One*, *10*(5), e0127317.

Morgan, M. S. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, *12*(2), 317–329.

Morgan, M. S. (2012). *The world in the model: How economists work and think*. Cambridge University Press.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science* (Vol. 52). Cambridge University Press.

Morgan, M. S., & Radder, H. (2003). Experiments without material intervention: Model experiments, virtual experiments and virtually experiments. *The Philosophy of Scientific Experimentation*.

Müller, B., Balbi, S., Buchmann, C. M., De Sousa, L., Dressler, G., Groeneveld, J., … Nolzen, H. (2014). Standardised and transparent model descriptions for agent-based models: Current status and prospects. *Environmental Modelling & Software*, *55*, 156–163.

Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., … Schwarz, N. (2013). Describing human decisions in agent-based models–ODD+ D, an extension of the ODD protocol. *Environmental Modelling & Software*, *48*, 37–48.

NASA Goddard Institute for Space Studies. (2021). GISS Surface Temperature Analysis (v4). Retrieved April 30, 2021, from https://data.giss.nasa.gov/gistemp/graphs_v4/

Neelin, J. D. (2010). *Climate change and climate modeling*. Cambridge University Press.

Network, C. (2020). Open Code Badge. Retrieved March 16, 2021, from https://www.comses.net/resources/open-code-badge/

O'Connor, C., & Weatherall, J. O. (2016). Black holes, black-scholes, and prairie voles: An essay review of simulation and similarity, by michael weisberg. University of Chicago Press Chicago, IL.

O'Sullivan, D., Evans, T., Manson, S., Metcalf, S., Ligmann-Zielinska, A., & Bone, C. (2016). Strategic directions for agent-based modeling: avoiding the YAAWN syndrome. *Journal of Land Use Science*, *11*(2), 177–187.

Odenbaugh, J. (2003). Complex systems, trade-offs, and theoretical population biology: Richard Levin's "strategy of model building in population biology" revisited. *Philosophy of Science*, *70*(5), 1496–1507.

Odenbaugh, J. (2006). The strategy of "The strategy of model building in population biology." *Biology and Philosophy*, *21*(5), 607–621.

Odenbaugh, J. (2015). Semblance or similarity? Reflections on Simulation and Similarity. *Biology & Philosophy*, *30*(2), 277–291. https://doi.org/10.1007/s10539-014-9446-y

Odenbaugh, J. (2018a). Building Trust, Removing Doubt? Robustness Analysis and Climate Modeling. In *Climate Modelling* (pp. 297–321). Springer.

Odenbaugh, J. (2018b). Models, models, models: a deflationary view. *Synthese*, 1–16.

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology and Philosophy*, *26*(5), 757–771. https://doi.org/10.1007/s10539-011-9278-y

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., & Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: a review. *Annals of the Association of American Geographers*, *93*(2), 314–337.

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science*, *11*(4), 349–368. https://doi.org/10.1007/s10699-005-3196-x

Parker, W. S. (2009a). Does Matter Really Matter? Computer Simulations, Experiments, and Materiality Wendy S. Parker. *Synthese*, *169*(3), 1–25.

Parker, W. S. (2009b). II—Confirmation and adequacy-for-purpose in climate modelling. In *Aristotelian Society Supplementary Volume* (Vol. 83, pp. 233–249). Wiley Online Library.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 263–272.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science*, *78*(4), 579–600.

Parker, W. S. (2013a). Computer simulation. *The Routledge Companion to Philosophy of Science*, *2*, 135–145.

Parker, W. S. (2013b). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(3), 213–223. https://doi.org/10.1002/wcc.220

Parker, W. S. (2014). Simulation and understanding in the study of weather and climate. *Perspectives on Science*, *22*(3), 336–356.

Parker, W. S. (2015). Getting (even more) serious about similarity. *Biology and Philosophy*, *30*(2), 267–276. https://doi.org/10.1007/s10539-013-9406-y

Parker, W. S. (2016). Reanalyses and observations: What's the difference? *Bulletin of the American Meteorological Society*, *97*(9), 1565–1572.

Parker, W. S. (2018). The significance of robust climate projections. In *Climate Modelling* (pp. 273–296). Springer.

Parker, W. S. (2020a). Evidence and Knowledge from Computer Simulation. *Erkenntnis*, 1–18.

Parker, W. S. (2020b). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457–477.

Parker, W. S., & Risbey, J. S. (2015). False precision, surprise and improved uncertainty assessment. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2055), 20140453.

Pascoe, C., Lawrence, B. N., Guilyardi, E., Juckes, M., & Taylor, K. E. (2019). Designing and documenting experiments in CMIP6. *Geosci. Model Dev. Discuss*, 1–27.

Pasini, A. (2005). *From observations to simulations: a conceptual introduction to weather and climate modelling*. World Scientific.

Pasini, A., Lorè, M., & Ameli, F. (2006). Neural network modelling for the analysis of

forcings/temperatures relationships at different scales in the climate system. *Ecological Modelling*, *191*(1), 58–67.

Pasini, A., & Mazzocchi, F. (2015). A multi-approach strategy in climate attribution studies: Is it possible to apply a robustness framework? *Environmental Science & Policy*, *50*, 191–199.

Pasini, A., Triacca, U., & Attanasio, A. (2012). Evidence of recent causal decoupling between solar radiation and global temperature. *Environmental Research Letters*, *7*(3), 34020.

Pershin, Y. V., La Fontaine, S., & Di Ventra, M. (2008). Memristive model of amoeba's learning. *Physical Review E*, *80*(2), 1–6. https://doi.org/10.1103/PhysRevE.80.021926

Peterson, T. C., Connolley, W. M., & Fleck, J. (2008). The myth of the 1970s global cooling scientific consensus. *Bulletin of the American Meteorological Society*, *89*(9), 1325–1338.

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., … Cinquini, L. (2020). Coordinating an operational data distribution network for CMIP6 data. *Geoscientific Model Development Discussions*, 1–22.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Programme, W. C. R. (2016). CMIP6.

Rasool, S. I., & Schneider, S. H. (1971). Atmospheric carbon dioxide and aerosols: Effects of large increases on global climate. *Science*, *173*(3992), 138–141.

Reid, C. R., & Beekman, M. (2013). Solving the Towers of Hanoi - how an amoeboid organism efficiently constructs transport networks. *The Journal of Experimental Biology*, *216*(9), 1546–1551. https://doi.org/10.1242/jeb.081158

Reid, C. R., Beekman, M., Latty, T., & Dussutour, A. (2013). Amoeboid organism uses extracellular secretions to make smart foraging decisions. *Behavioral Ecology*, *24*(4), 812–818. https://doi.org/10.1093/beheco/art032

Reid, C. R., Latty, T., Dussutour, A., & Beekman, M. (2012). Slime mold uses an externalized spatial "memory" to navigate in complex environments. *Proceedings of the National Academy of Sciences*, *109*(43), 17490–17494. https://doi.org/10.1073/pnas.1215037109

Rice, C. C. (2016). Factive scientific understanding without accurate representation. *Biology & Philosophy*, *31*(1), 81–102.

Rice, C. C. (2019). Understanding realism. *Synthese*, 1–25.

Richiardi, M. G. (2017). The future of agent-based modeling. *Eastern Economic Journal*, *43*(2), 271–287.

Rohrlich, F. (1990). Computer simulation in the physical sciences. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1990, pp. 507–518). Philosophy of Science Association.

Rosinski, J. M., & Williamson, D. L. (1997). The accumulation of rounding errors and port validation for global atmospheric models. *SIAM Journal on Scientific Computing*, *18*(2), 552–564.

Ross, L. (2021). The truth about better understanding? *Erkenntnis*, 1–24.

Sachse, R., Westermeier, A., Mylo, M., Nadasdi, J., Bischoff, M., Speck, T., & Poppinga, S. (2020). Snapping mechanics of the Venus flytrap (Dionaea muscipula). *Proceedings of the National Academy of Sciences*, *117*(27), 16035–16042.

Saigusa, T. (2008). Amoebae Anticipate Periodic Events. *Physical Review Letters*, *018101*(JANUARY), 1–4. https://doi.org/10.1103/PhysRevLett.100.018101

Sansores, C., Pavón, J., & Gómez-Sanz, J. (2005). Visual modeling for complex agent-based simulation systems. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation* (pp. 174–189). Springer.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24.

Schneider, S. H., & Dickinson, R. E. (1974). Climate modeling. *Reviews of Geophysics*, *12*(3), 447–493.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3.

Schönwiese, C.-D., Walter, A., & Brinckmann, S. (2010). Statistical assessments of anthropogenic and natural global climate forcing. An update. *Meteorologische Zeitschrift*, *19*(1), 3–10.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, *69*(1), 275–300.

Shukla, J., Palmer, T. N., Hagedorn, R., Hoskins, B., Kinter, J., Marotzke, J., … Slingo, J. (2010). Toward a new generation of world climate research and computing facilities. *Bulletin of the American Meteorological Society*, *91*(10), 1407–1412.

Sismondo, S. (1999). Models, simulations, and their objects. *Science in Context*, *12*(2), 247–260.

Skillings, D. J. (2015). Mechanistic Explanation of Biological Processes. *Philosophy of Science*, *82*(5), 1139–1151.

Skipper Jr, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: natural selection. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 327–347.

Smaldino, P. E., Calanchini, J., & Pickett, C. L. (2015). Theory development with agent-based models. *Organizational Psychology Review*, *5*(4), 300–317.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104.

Smith, J. M. (1964). Group selection and kin selection. *Nature*, *201*(4924), 1145.

Squazzoni, F. (2010). The impact of agent-based models in the social sciences after 15 years of

incursions. *History of Economic Ideas*, 197–233.

Stocker, T. (2014). *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press. https://doi.org/10.1080/02698590903007212

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515.

Sun, Z., Di, L., Cash, B., & Gaigalas, J. (2020). Advanced cyberinfrastructure for intercomparison and validation of climate models. *Environmental Modelling & Software*, *123*, 104559.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075.

Tero, A., Takagi, S., Ito, K., Bebber, D. P., Fricker, M. D., Yumiki, K., … Nakagaki, T. (2010). Rules for Biologically Inspired Adaptive Network Design. *Science*, *327*(January), 439–442. https://doi.org/citeulike-article-id:6578033

Touzé-Peiffer, L., Barberousse, A., & Le Treut, H. (2020). The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(4), e648.

Trout, J. D. (2016). *Wondrous Truths: The Improbable Triumph of Modern Science*. Oxford University Press.

Turrell, A. (2016). Agent-based models: understanding the economy from the bottom up. *Bank of England Quarterly Bulletin*, Q4.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *Journal of Philosophy*, *92*(2), 345–381. https://doi.org/10.2307/2026571

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628.

Van Gelder, T., & Port, R. F. (1995). It's about time: An overview of the dynamical approach to cognition. *Mind as Motion: Explorations in the Dynamics of Cognition*, *1*, 43.

van Marle, M. J. E., Kloster, S., Magi, B. I., Marlon, J. R., Daniau, A. L., Field, R. D., … Kehrwald, N. M. (2017). Historic global biomass burning emissions for CMIP6 (BB4CMIP) based on merging satellite observations with proxies and fire models (1750–2015), Geosci. Model Dev., 10, 3329–3357.

Veit, W. (2020). Model Anarchism.

Verdes, P. F. (2007). Global warming is driven by anthropogenic emissions: a time series analysis approach. *Physical Review Letters*, *99*(4), 48501.

Villamor, G. B., Troitzsch, K. G., & Van Noordwijk, M. (2013). Validating human decision making in an agent-based land-use model. In *Proceedings of the 20th International Congress on Modelling and Simulation (MODSIM): Adapting to Change: The Multiple Roles of Modelling, Adelaide, Australia* (Vol. 16). Citeseer.

Vogel, D., Nicolis, S. C., Perez-Escudero, A., Nanjundiah, V., Sumpter, D. J. T., & Dussutour, A. (2015). Phenotypic variability in unicellular organisms: from calcium signalling to social behaviour. In *Proceedings. Biological sciences / The Royal Society* (Vol. 282, pp. 20152322-). The Royal Society. https://doi.org/10.1098/rspb.2015.2322

Wallentin, G. (2017). Spatial simulation: a spatial perspective on individual-based ecology—a review. *Ecological Modelling*, *350*, 30–41.

Walmsley, L. D. (2020). The strategy of model building in climate science. *Synthese*, 1–21.

Washington, W. M., & Parkinson, C. (2005). *Introduction to three-dimensional climate modeling*. University science books.

Watt, K. E. F. (1956). The choice and solution of mathematical models for predicting and maximizing the yield of a fishery. *Journal of the Fisheries Board of Canada*, *13*(5), 613–645.

Watt, K. E. F. (1962). Use of mathematics in population ecology. *Annual Review of Entomology*, *7*(1), 243–260.

Watt, K. E. F., & Watt, K. E. F. (1968). *Ecology and resource management; a quantitative approach*. McGraw-Hill,.

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., … Kay, J. E. (2017). The cloud feedback model intercomparison project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384.

Weisberg, M. (2004). Qualitative theory and chemical explanation. *Philosophy of Science*, *71*(5), 1071–1081.

Weisberg, M. (2006a). Forty Years of "The Strategy": Levins on Model Building and Idealization. *Biology and Philosophy*, *21*(5), 623–645.

Weisberg, M. (2006b). Robustness Analysis. *Philosophy of Science*, *73*(5), 730–742. https://doi.org/10.1007/s001900100162

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press. Retrieved from http://books.google.com/books?id=rDu5e532mIoC&pgis=1

Weisberg, M. (2014). Understanding the Emergence of Population Behavior in Individual-Based Models. *Philosophy of Science*, *81*(5), 785–797.

Weisberg, M. (2015). Biology and philosophy symposium on simulation and similarity: Using models to understand the world. *Biology & Philosophy*, *30*(2), 299–310.

Weisberg, M., & Reisman, K. (2008). The Robust Volterra Principle. *Philosophy of Science*, *75*(1), 106–131. https://doi.org/10.1086/588395

Weiskopf, D. (2017). The explanatory autonomy of cognitive models. *Integrating Psychology and Neuroscience: Prospects and Problems. Oxford: Oxford University Press, Forthcoming*.

Wiggins, B. J., & Chrisopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202.

Wilensky, U., & Rand, W. (2007). Making models match: Replicating an agent-based model. *Journal of Artificial Societies and Social Simulation*, *10*(4), 2.

Wilensky, U., & Rand, W. (2015). *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT Press.

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, *190*(6), 997–1016.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. *Scientific Inquiry and the Social Sciences*, 124–163.

Wimsatt, W. C. (1986). Forms of aggregativity. In *Human nature and natural knowledge* (pp. 259–291). Springer.

Wimsatt, W. C. (1987). False models as means to truer theories. *Neutral Models in Biology*, 23–55.

Wimsatt, W. C. (2006). Aggregate, composed, and evolved systems: Reductionistic heuristics as means to more holistic theories. *Biology and Philosophy*, *21*(5), 667–702.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models:

Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, *10*(2), 8.

Winsberg, E. (1999). Sanctioning Models: The Epistemology of Simulation. *Science in Context*, *12*(02), 275–292. https://doi.org/10.1017/S0269889700003422

Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representations. *Philosophy of Science*, *68*(S3), S442–S454.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, *70*(1), 105–125.

Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.

Winsberg, E. (2018a). *Philosophy and climate science*. Cambridge University Press.

Winsberg, E. (2018b). What does robustness teach us in climate science: a re-appraisal. *Synthese*, 1–24.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*(3), 393–472.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197–254.

Woodward, J. (2003). Making Things Happen: A Theory of Causal Explanation. Oxford University Press. https://doi.org/10.1080/10705510709336742

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240.

Woodward, J. (2011). Data and phenomena: a restatement and defense. *Synthese*, *182*(1), 165–179.

Zabusky, N. J. (1987). Grappling with complexity. *Physics Today*, *40*(10), 25–27.

Zhang, X., Adamatzky, A., Chan, F. T. S., Deng, Y., Yang, H., Yang, X.-S., … Mahadevan, S. (2015). A biologically inspired network design model. *Scientific Reports*, *5*, 10794. https://doi.org/10.1038/srep10794