EXPERIMENTS, SIMULATIONS, AND LESSONS FROM EXPERIMENTAL EVOLUTION

Emily C. Parke

A DISSERTATION

in

Philosophy

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

_____

Michael Weisberg, Associate Professor of Philosophy

Graduate Group Chairperson

_____

Michael Weisberg, Associate Professor of Philosophy

Dissertation Committee

Karen Detlefsen, Associate Professor of Philosophy and Education
Daniel Singer, Assistant Professor of Philosophy
Paul Sniegowski, Professor of Biology

EXPERIMENTS, SIMULATIONS, AND LESSONS FROM EXPERIMENTAL EVOLUTION

ACKNOWLEDGMENT

I am immensely grateful to my committee members, all of whom have inspired, challenged, and encouraged me not only in writing this dissertation but throughout my years as a graduate student. Michael Weisberg has been a great source of insight and support, always encouraging me to develop my own projects while guiding me to take them in fruitful directions and offering knowledgeable feedback every step of the way. I have benefited from all of his advice over the years, on everything from how to give talks to which font to use, and feel very lucky to have him as a supervisor. I am thankful to Karen Detlefsen for her support and generosity with her time, wealth of knowledge on the history and philosophy of the life sciences, and meticulous feedback on my writing. I am grateful to Paul Sniegowski for his thoughtful mentorship, inspiring conversations about science and philosophy and teaching, and for welcoming me into his lab and patiently teaching me so much about biology over the years. I was fortunate to have Dan Singer join my committee in my last two years at Penn, and have appreciated his careful comments on my work which have helped improve my attention to detail, his knowledge and perspective on formal matters, and general encouragement as I finished the PhD.

While not on my official committee, I am also particularly indebted to three other mentors: Mark Bedau, for his ongoing advising, encouragement, and comments on everything I've written in the last 11+ years; Zoltan Domotor, who has been a valuable resource on all matters in philosophy of science; and Mary Morgan, whose work has been an important source of inspiration, for many great discussions and feedback on my work. I am grateful to all of my past and present colleagues in philosophy at Penn, in particular the members of Science Club, and to my fellow members of the Sniegowski Lab for being so welcoming and teaching me everything I know about experimental science: Matthew Gazzara, Angela Halstead, Meredith Hyun, Dave Katzianer, Kate Kerpen, Brooks Martino, Eugene Raynes, Kathleen Sprouffske, Ben Sprung, and especially Chris Gentile, Mitra Eghbal, and Tanya Singh.

I have also benefitted from feedback (both written and in conversations) on portions of the material in this dissertation from a number of others: Wes Anderson, David Barack, Matt Bateman, Brett Calcott, Mark Colyvan, Lindley Darden, Alkistis Elliott-Graves, Karen Kovaca, Peter Godfrey-Smith, Ben Kerr, J. J. LaTourelle, Alan Love, Margaret Morrison, Valerie Racine, Eric Saidel, Carlos Santana, Kyle Stanford, Kim Sterelny, and Michael Travisano.

Finally thanks to my husband Brett, who is a constant source of support, inspiration, and good conversations about biology and philosophy and everything else.

ABSTRACT

EXPERIMENTS, SIMULATIONS, AND LESSONS FROM EXPERIMENTAL EVOLUTION

Emily C. Parke

Michael Weisberg

Philosophers and scientists have sought to draw methodological distinctions among different kinds of experiments, and between experimentation and other scientific methodologies. This dissertation focuses on two such cases: hypothesis-testing versus exploratory experiments, and experiment versus simulation. I draw on examples from experimental evolution—evolving organisms in a controlled laboratory setting to study evolution via natural selection in real time—to challenge the way we think about these distinctions. In the case of hypothesis-testing versus exploratory experiments, philosophers have distinguished these categories in terms of the role of theory in experiment. I discuss examples from experimental evolution which occupy the poorly characterized middle ground between the two categories. I argue that we should take more seriously the point that multiple theoretical backgrounds can come into play at multiple points in an experiment, and propose some new contributions toward clarifying the conceptual space of experimental inquiry. In the case of experiment versus simulation, people have attempted to clearly delineate cases of science into these two categories, and base judgments about their epistemic value on these categorizations. I discuss and reject two arguments for the epistemic superiority of experiments over simulations: (1) Experiments put scientists in a better position to make valid inferences about the natural world; (2) Experiments are a superior source of surprises or novel insights. Both of these claims are false as generalizations across science. Focusing on the experiment/simulation

distinction as a basis for in-principle judgments about epistemic value focuses us on the wrong issues. This leaves us with a question: What should we focus on instead? I offer preliminary considerations for a framework for evaluating inferences from objects of study to targets of inquiry in the world, which departs from the problematic custom of basing such evaluations on questions like "Was it an experiment or a simulation?" This framework is based on the ideas of capturing relevant similarities while appropriately accounting for what researchers already know and what they are trying to learn by asking the scientific question at hand.

TABLE OF CONTENTS

1. Experiments **1**

1.1. What is an Experiment? 1

1.2. Different Kinds of Experiments 8

1.3. Overview of the Dissertation 10

2. Experimental Evolution **13**

2.1. Lenski's Long-Term Evolution Experiment 16

2.1.1. High Mutation Rates *21*

2.1.2. Punctuated Evolution *22*

2.1.3. Citrate Utilization *24*

2.1.4. Laboratory Natural Selection? *25*

2.1.5. The Lenski Experiment? *27*

2.2. LNS Experiments Beyond Lenski: History and Current Directions 28

2.3. Tapes of Life, Darwin's Mistake, and Other Challenges 33

3. Hypothesis-Testing and Exploratory Experiments **37**

3.1. Hypothesis Testing: The Classic View of Experiments 39

3.2. Exploratory Experiments: A Response to the Classic View 42

3.3. Spaces of Experimental Inquiry 52

3.4. Theories Can Play Many Roles in Experiments 60

4. Experiments and Simulations **68**

4.1. Some Preliminary Points About Experiments and Simulations 69

4.1.1. Simulations or Models? *72*

4.2. The Argument about Inferential Power 75

4.3. Responses to the Materiality Thesis 77

4.3.1. Experiments Are Not Necessarily Designed To Be Like Particular Targets *78*

4.3.2. Material Correspondence: Hard to Evaluate, Not Always the Goal *81*

4.4. Surprise 89

4.4.1. Response to the Surprise Claim *92*

4.5. Experimenters Almost Never Study Their Targets Directly 103

4.6. What's the Difference between Experiment and Simulation? 107

5. Conclusion: Evaluating Inferences about the Natural World **111**

6. Bibliography **130**

LIST OF ILLUSTRATIONS

Figure 1: An illustration of the serial transfer protocol used in Lenski's experiment.....17

Figure 2: A TA plate showing red and white Ara^{-}/Ara^{+} colonies (photo credit: Paul Sniegowski).....19

Figure 3: Cell size as a function of time (over 3,000 generations) in one of the 12 populations on the Lenski experiment. From (Elena, Cooper, & Lenski 1996, Figure 1)...23

Figure 4: Dallinger's drawing of the apparatus used in his evolution experiment (from Dallinger 1887).....30

Figure 5: Mutation rates of wild type, single and double mutator *E. coli* (from (Gentile 2012); see also (Gentile et al. 2011)). Estimates are of the per base pair, per generation genomic mutation rates with 95% confidence intervals for wild type *E. coli*, single mutators with the *mutL13* allele (which confers deficiency in mismatch repair), and double mutators with *mutL13* and *dnaQ905* alleles (the latter confers deficiency in DNA proofreading). Single mutators have a genomic mutation rate 100–fold higher than the wild type; double mutators have a genomic mutation rate 45–fold higher than single mutators.....94

Figure 6: (a) A representative grid of cells in the Game of Life. (b) A glider gun, with a stream of gliders moving off toward the lower right.....99

1. Experiments

Experiments play a central role in scientific practice. Throughout the history of philosophy of science they have been largely neglected in favor of focusing attention on theories and models. Until several decades ago philosophers did not say much about experiments, and when they did, they talked about them mainly just as means to the end of linking theories and models to the world. Philosophy of scientific experimentation is still a surprisingly small but growing area in philosophy of science. Two recent trends in this area have sought to categorize different kinds of experiments, and to distinguish experiment from other scientific methodologies like simulation and observation. A key motivation for carving these sorts of methodological lines is to argue that they correspond to differences in epistemic value. For example, scientists as well as philosophers commonly claim that experiments have a privileged status over simulations because they put researchers in a better position to generate trustworthy inferences about the natural world. In this dissertation I question the sharpness of some methodological distinctions that have been drawn regarding experiments, and the validity of using methodological categories—experiment versus simulation, in particular—as markers of epistemic differences. I propose a framework for evaluating inferences about the natural world which departs from the problematic custom of basing such evaluations on questions like "Was it an experiment or a simulation?"

1.1. What is an Experiment?

The key characteristic feature of scientific experimentation is intervention. Here are two examples:

A classic case, often featured in textbook introductions in discussions of the scientific method, is Francesco Redi's series of experiments on insect generation. Redi is the seventeenth-century natural philosopher credited with initiating the end of belief in spontaneous generation. Since antiquity people had believed that inanimate matter could generate new life: riverbeds could give birth to eels (Aristotle, *History of Animals* 569a10-25; 570a4-12) and grain or sweaty piles of clothing left in a barn could give birth to mice (Fry 2000). In Redi's experiments on insect generation he set out to challenge the claim that nonliving matter can generate new life (Redi 1668). He placed samples of organic matter in jars, left some uncovered and sealed others with paper and twine, and reported that by preventing insects from gaining access to the latter jars' contents, no larvae would appear therein. This demonstrated that the life forms which were commonly observed "spontaneously" appearing on meat and other organic matter got there because flies had laid their eggs there, not because of some intrinsic life-generating property in the meat itself. He used hundreds of different kinds of plant and animal matter to demonstrate this point, including flowers, grasses, and an impressive array of meats ranging from dog and eel to lion and water buffalo. In Redi's detailed descriptions of thousands of repetitions of this experiment he continually emphasizes the necessity of many iterated rounds of gathering evidence to confirm his beliefs, and the careful use of what we now call controls: experimental setups which isolate single independent variables (for further discussion of Redi's work see Parke 2014a).

Another example, which has been called the "most beautiful experiment in biology" (Holmes 2001), is Meselsohn and Stahl's 1957 demonstration of the mechanism of DNA replication (see discussion in Weber 2014). In particular, their experiment confirmed Watson and Crick's prediction that DNA replication is semiconservative: When

molecules are copied, one strand of each resulting double helix is newly synthesized, while the other strand remains from the unwound "parent" molecule. Meselsohn and Stahl grew bacteria for a number of generations in a growth medium in which the only available nitrogen was a heavy isotope, ^{15}N, which resulted in the bacteria's DNA molecules containing only heavy nitrogen. They then transferred the bacteria to growth medium containing only standard lighter nitrogen, ^{14}N, and extracted their DNA after one generation and after two generations in the new growth medium. They put the extracted DNA in a highly sensitive measuring device, capable of separating DNA containing heavy nitrogen, light nitrogen, or a hybrid intermediate. They found that after one round of DNA replication the cells' DNA contained the hybrid intermediate, and after two rounds the lighter nitrogen version reappeared. This confirmed the hypothesis that DNA replication is semiconservative.

These are both classic examples where experimenters designed interventions to test hypotheses. Until the last several decades, philosophers of science talked about experiments almost exclusively as means to the end of testing hypotheses. In *Philosophy of Natural Science*, Hempel talks about experiments in terms of "bringing about the conditions *C* and checking whether *E* occurs as implied by the hypothesis," where the *C* and *E* referred to are terms in a hypothesis of the form: If conditions of kind *C* occur, then an event of kind *E* will occur (Hempel 1966, p. 20). The examples discussed above fit this sort of model nicely. In the Redi case, the hypothesis in question could be stated as "If jars are sealed to prevent flies from entering them, then no maggots will appear on their contents." In the Meselsohn-Stahl case, it could be stated as "If the semiconservative model of DNA replication is correct, then we should expect to see a certain pattern of

DNA nitrogen content when moving bacteria from growth medium with ^{15}N to medium with ^{14}N."

The last few decades have seen philosophers and historians of science increasingly giving experiments more attention, and paying attention to the fact that they do not all fit the standard hypothesis-testing model. Pioneering works in this area include Ian Hacking's book *Representing and Intervening* (1983), Allan Franklin's *The Neglect of Experiment* (1989), and Peter Galison's *How Experiments End* (1987). These works laid the early groundwork for a growing literature on experiments and their role in science. There is plenty to say about the contributions of this early work, but the key headlines are: taking experiments more seriously as central to understanding scientific practice; acknowledging that they often play a more complex, exploratory role than mere means to the end of testing hypotheses; beginning to account for the many ways scientists can engage methodologically and epistemically with their experimental objects of study; and exploring the ways that experiments overlap and differ from other scientific methodologies like simulation, observation, and measurement.

I will understand experiments throughout the dissertation as interventions in physical systems, consisting of physical entities and their environments in the laboratory or out in nature. The physical systems experimenters study vary widely, from meat in jars to bacteria in test tubes to electrons in particle accelerators to trees in forests. Okasha nicely characterizes the intuitive idea of the kind of intervention which takes place in experiments (contrasting experiment with observation):

> Experiments involve actively intervening in the course of nature, as opposed to observing events that would have happened anyway. When a molecular biologist inserts viral DNA into a bacterium in his laboratory, this is an experiment; but when an astronomer points his telescope at the heavens, this is an observation. Without the biologist's handiwork the bacterium would never have contained foreign DNA; but the planets would have

> continued orbiting the sun whether or not the astronomer had directed his telescope skyward. (Okasha 2011, p. 222)

Following Woodward (2003, 2013) and others, I take the notion of an ideal experimental intervention to have to do with "surgical" change in a variable. As he puts it, regarding two variables *A* and *B*, such changes are "of such a character that if any change occurs in *B*, it occurs only as a result of its causal connection, if any, to *A* and not in any other way. In other words, the change in *B*, if any, that is produced by the manipulation of *A* should be produced only via a causal route that goes through *A*" (Woodward 2013).

People have also characterized the core methodology of experiments as involving putting a system in a state. Peschard (2012, p. 2) exemplifies this when she defines an experiment as follows:

> … experimentation on a physical system S can be thought of as a procedure that consists in, at least:
>
> 1. Preparing the system S in a certain state, by fixing initial and boundary conditions, and selectively putting under control the parameters that have an effect on the outcomes of measurement, the *active parameters*.
>
> 2. Letting the system evolve. The evolution of S is characterized by the evolution of a set of physical quantities characterizing the state of S, the *state variables*.
>
> 3. Recording the evolution of S through a sequence of states when the values of some of these parameters are varied; analyzing the results.

Along similar lines, Winsberg (2009, p. 582) says that when one experiments, one "intervenes in [a system] by putting it in a particular initial state and observes its subsequent states to learn about its properties in light of that intervention." It might seem like there are two views of experiment at work here, intervention as outlined above versus putting systems in states. I think the latter is just one way to talk about intervention. Talking about it that way implies a certain kind of successful directed intervention. When we put a system in a state we are always intervening, but we can intervene without putting

a system in a state (understood as intending to, and successfully, putting it in a particular state). Parker (2009, p. 487) nicely ties together these two ways to talk about experiments in her characterization:

> An *experiment* can be characterized as an investigative activity that involves intervening on a system in order to see how properties of interest of the system change, if at all, in light of that intervention. An *intervention* is, roughly, an action intended to put a system into a particular state, and that does put the system into a particular state, though perhaps not the one intended.

So I am inclined to say that the intervention and putting-systems-in-states views are not two different views of what an experiment is, but rather two ways to talk about what experimenters can be up to when they intervene.

Understanding experiments as interventions in physical systems distinguishes them from computer simulations, which are studies of computational models rather than physical systems. This position is somewhat controversial, because some people want to say that computer simulations *are* experiments. I discuss these views in Chapter 4, but do not want to take a strong position in that debate here or there. A key point I make in that discussion is that we should not hang so much on whether we label cases of scientific inquiry as experiments or simulations. A lot of ink has been spilled arguing for various ways to neatly carve up cases of research into these methodological categories, for the purposes of having categories to base judgments on (for instance, judgments about the epistemic privilege of experiments over simulations). I will argue that such categorization projects focus us on the wrong issues. The relevant points for evaluating cases of scientific inquiry are context-sensitive issues regarding the connection between the object of study and target of inquiry in question, how much a researcher knows about each, and what sort of scientific question she is asking—not whether we have labeled the case as an experiment or a simulation or something else, per se. My reason for understanding

experiments as interventions in material systems is not to definitively deny simulations status as experiments, but to maintain what I think is an important methodological distinction between studying physical systems and studying computational or mathematical models on computers. I say more about this methodological distinction in Chapters 4 and 5.

I should underline at the outset that it is not a goal of this dissertation to argue that methodological categorizations are useless or meaningless, or that we should stop using terms like 'experiment' and 'simulation' to refer to research activities. 'Experiment' does and should refer to a way of doing science. My objections will be to views that make generalized claims about science like "Experiment is epistemically better than observation because it involves active manipulation," or "Experiments allow us to make better inferences about the natural world than simulations." Objecting to such claims is compatible with endorsing the view that there are meaningful methodological differences between experiment, simulation, and observation. I am not objecting to the entire project of distinguishing them.

This is probably as good a place as any to clarify some key terminology regarding experiments which I will use throughout the dissertation:

- **Object**, or object of study: The system a scientist engages with in order to learn about her target of inquiry. In an experiment, the object is an experimental system. In a simulation, the object is a model.
- **Experimental system:** The complex of entities and processes (in most biology experiments these are organisms, their environment, and an experimental protocol) which together comprise an experimentalist's object.

- **Target**,[1] or target of inquiry: The system a scientist ultimately aims to generate knowledge about through studying or intervening in her object. Most often, but not always, the target is some system in the natural world. Very rarely, the object and target are identical. Almost always, they are not (I discuss this point further in Chapter 4).

1.2. Different Kinds of Experiments

Chapter 3 will focus on two kinds of experiments, hypothesis-testing versus exploratory experiments, distinguished according to their relationships with theory. There is another way kinds of experiments are commonly distinguished which is worth mentioning: laboratory versus field versus natural experiments. Laboratory experiments are distinguished from field and natural experiments according to where they take place and the degree of control researchers have. The former take place, of course, in the lab, and the latter two take place in the world outside the lab. Laboratory experiments generally afford a greater degree of control than field and natural experiments because researchers have control, at least in principle, over far more details of the entities and environments constituting their experimental systems.

Field experiments are distinguished from laboratory experiments primarily according to their location: they take place out in nature rather than in the confines of a laboratory. Unlike laboratory experiments where researchers have control (at least in principle) over many variables, field experiments tend to involve manipulating a single

[1] I am using the terms 'object' and 'target' in this way following Winsberg (2009) and others. It is important to clarify this usage because people have used the word 'target' in different ways: usually to refer to what I call the target here, but occasionally to refer to what I call the (experimental) object here (Peschard 2012; Winsberg 2010, p. 31, p. 35), which can lead to confusion.

variable, like removing or blocking a species from a particular geographic location, or introducing a new species to a localized environment like an island (Diamond 1983; Irschick & Reznick 2009).

The name 'natural experiments' is somewhat misleading as a way to distinguish a separate category from field experiments, because both occur out in nature rather than in the confines of the laboratory. In the case of natural experiments, 'natural' refers not (just) to the setting but to the way that the experimental intervention comes about. In laboratory and field experiments, the experimenter himself is responsible for the intervention; in a natural experiment, he is not. This can be either because somebody else is or nobody is. Woodward defines natural experiments as those which "typically [involve] the occurrence of processes in nature that have the characteristics of an intervention but do not involve human action or at least are not brought about by deliberate human design" (2003, p. 94). A classic example of a natural experiment is the case of the cholera outbreak in London in the 1950s. A physician named John Snow identified a particular public water pump as the source of the outbreak, and found strong associations between using water from that supply and becoming ill or dying of cholera. He also found that districts of the city whose water came from a different supplier had lower rates of cholera. His study of the outbreak in London involved a setup very much like the interventions experimenters intentionally cause in the lab or the field: One group used one water supply, another group used another, and this setup presented an opportunity to test the claim that using water from the former source causes cholera outbreaks. Only in cases like this natural experiment, nobody intentionally caused the intervention. Other examples of natural experiments include cases where species are introduced to novel habitats (either intentionally by humans or not), which afford researchers an unplanned "natural"

experimental setup for studying the evolutionary changes which result from these introductions (see discussion in Irschick & Reznick 2009).

1.3. Overview of the Dissertation

The rest of this dissertation addresses some methodological distinctions that have been made regarding experiments, both within the context of experimentation and between experimentation and and simulation. Throughout the dissertation I rely on examples from a research area which provides particularly provocative cases for challenging the methodological distinctions in question: experimental evolution. Chapter 2 gives an overview of experimental evolution, which involves researchers propagating populations of organisms in the laboratory as a means to study evolution via natural selection in real time. This research area has received little attention from philosophers of science; it deserves more attention, as it offers a wealth of interesting cases to think about for philosophers interested in topics in scientific methodology, evolution, and ecology. I explain the history and scope of experimental evolution and describe in detail a noteworthy example, Richard Lenski's long-term evolution experiment. I return to the long-term evolution experiment throughout the rest of the dissertation in my discussion of methodological categorizations and how scientists make inferences from objects to targets.

I then turn to the distinctions which have been drawn among different kinds of experiments. Section 1.2 briefly discussed kinds of experiments distinguished by the methodological issues of where they take place and the extent to which the researcher herself is responsible for the intervention. Chapter 3 focuses on another way that kinds of

experiments are distinguished: according to the role (or better, I argue, roles) played by theory. The classic view of experiments discussed earlier in this chapter says that theory comes first, and the point of experiments is to test the hypotheses that theory generates. Recent literature has rightly pointed out that this is not the whole picture; experiments can also be exploratory. Chapter 3 addresses classic hypothesis-testing versus exploratory experiments, which are methodologically more open-ended and have as their aim filling gaps in theory or generating entirely new theory, rather than testing specific hypotheses. People have had a lot to say about hypothesis-testing and about exploratory experiments, and not much to say about the middle ground. I outline a preliminary account of how to fill in some aspects of this conceptual middle ground, drawing on examples of cases from experimental evolution which occupy it.

In Chapter 4 I turn to the distinction between experiment and simulation. Philosophers have argued that we can divide up cases of science into these two methodological categories, and that doing so gives us a basis for making epistemic judgments. These judgments invariably come out privileging experiments over simulations. I discuss two arguments that have been made for the superiority of experiments: (1) experiments put researchers in a better position than simulations do for making trustworthy inferences about the natural world, and (2) compared to simulations, experiments are a superior form of surprises or novel insights. I argue that both of these claims are false as generalizations across science. They are not false because experiments and simulations are epistemically on a par, or because simulations are in fact superior—rather, I argue, these methodological categorization-based assessments of science focus us on the wrong issues. We should stop looking to the experiment/simulation distinction to tell us anything in principle about epistemic value. Any judgments about the superiority of

experiments, or simulations, must be made in a context-sensitive way: in the context of a particular field or area of inquiry and the relevant shared background knowledge and research traditions.

This leaves us with a question: If categorical distinctions between experiment and simulation are a poor basis for judgments about inferential power, what is the right basis? Chapter 5 lays out initial considerations for a framework for evaluating inferences from scientific objects of study to targets of inquiry in the world, which departs from the problematic basis of beginning such evaluations by asking questions like "Was it an experiment or a simulation?" My account is based on the ideas of (1) capturing relevant similarities between object and target, and (2) appropriately accounting for what we already know and what we are trying to learn by asking the scientific question at hand. A more developed account is a topic for future work.

2. Experimental Evolution

It used to be the case that biologists' only reliable sources of empirical information about long-term evolution were the fossil record and living natural populations. There are limitations to using the fossil record to study evolution in action, particularly microevolution, because the evidence contained therein is patchy; it gives us only glimpses of the forms of living things at sporadic points in the past. Data from living natural populations, while in principle much richer in quantity and level of detail, similarly give us only snapshots. With rare exceptions, like the Grants' multi-decade study of Galapagos finch populations (Grant & Grant 2002), studies of natural populations do not give us dynamic data about evolutionary lineages through time. They give us static data about a population's current state and its gene pool, and we must then make inferences about how selection or other evolutionary forces have acted on that population.[2]

In the past few decades, a relatively new research area has added a richer source of long-term evolutionary data to the picture: experimental evolution, which involves propagating populations of organisms in a controlled laboratory setting to study evolution in real time. While experimental evolution actually dates back to the mid-nineteenth century (more on this below), it has become an increasingly active research area in biology just in recent years. The first major anthology on experimental evolution was published in 2009 (Garland & Rose 2009), a 2013 special issue of *Biology Letters* was devoted to current directions in the field (Bataillon, Joyce, & Sniegowski 2013 and papers

[2] This distinction between dynamic and static data is from Forber (2009).

therein), and the American Society for Microbiology just held its first conference on experimental evolution in summer 2014.

In the introduction to their recent anthology, Garland and Rose define experimental evolution as "research in which populations are studied across multiple generations under defined and reproducible conditions, whether in the laboratory or in nature," involving most, if not all, of the following features: "maintenance of control populations, simultaneous replication, observation over multiple generations, and the prospect of detailed genetic analysis" (Rose & Garland 2009, pp. 6–7). This broad definition includes natural experiments like studies of invasive species in the wild. It also includes artificial selection experiments and studies of domestication processes.

The discussion in this dissertation will focus on a particular way to do experimental evolution, a subset of what is covered by Garland and Rose's broad definition of the field: *laboratory natural selection* (henceforth LNS) experiments. LNS experiments are more interesting for the present purposes than some of the other kinds of research included under Garland and Rose's broad understanding of experimental evolution, like artificial selection and studies of natural populations. First, artificial selection has been happening, and has been talked about, for a long time; LNS experiments are newer and hence less discussed. LNS is importantly different from artificial selection: The populations are evolving via natural selection; researchers do not choose which individuals will carry on to the next generation based on any particular trait(s) they possess. As Roff and Fairbairn point out, "The major advantage of [LNS] over artificial selection is that the organisms are allowed to evolve relatively naturally in response to diverse selection acting on the whole phenotype, and hence the observed evolutionary processes may more closely mimic those that occur in nature" (Roff & Fairbarin 2009, p.

32). LNS is also importantly different from studies of natural populations: The controlled laboratory setting allows for much more direct, comprehensive analysis of the populations' genetic makeup and evolutionary history. As examples I discuss below illustrate, researchers much greater access to information about the organisms, their environments, their histories, and their genomes than they do in studies of natural populations..

The second, more important reason why I will focus on LNS experiments is that they present particularly compelling challenges for the boundaries philosophers of science have drawn regarding experiments. This is thanks to the combination of three particular features mentioned above: the degree of control afforded by the laboratory setting, the open-ended ability to watch evolution as it happens and collect extensive data throughout the process, and the "closeness" of the evolutionary processes in the lab to those occurring in nature. LNS experiments have features we typically think of as paradigm of non-experimental methodologies like simulation and observational studies. I return to this point at the end of the chapter.

Section 2.1 gives an overview of an iconic LNS experiment, which I return to as a core example throughout the dissertation: Richard Lenski's long-term evolution experiment. Section 2.2 goes into more detail about the history and scope of LNS experiments in general; Section 2.3 discusses some challenges this research area presents for ideas about the pace of evolutionary change and our ability to empirically test questions about the roles of history and contingency in evolution.

2.1. Lenski's Long-Term Evolution Experiment

The organisms of choice for experimental evolution are commonly microbes like bacteria. Bacteria are ideal subjects for a number of reasons: They have short generation times (from several to tens of generations per day), large populations of them can be stored in compact spaces, and they can be frozen and revived, allowing for easy comparisons and competitions between evolved populations and their ancestors (Forde & Jessup 2009; Gentile et al. 2011; Lenski et al. 1991; Travisano et al. 1995).

In February 1988, Richard Lenski's lab used a single ancestral genome of *Escherichia coli* to found twelve genetically identical[3] populations in twelve identical environments: flasks filled with 10 ml of bacterial growth medium, a nutrient broth containing glucose as a limiting resource. These *E. coli* were engineered to be incapable of recombination, that is, they reproduce completely asexually. Lenski's group was interested in observing the evolutionary dynamics of diversification and adaptation as they happened in the laboratory. The basic protocol of their experiment involves a simple serial transfer protocol: As the populations grow over the course of every 24 hours, they deplete the resources in their flasks, so every day a 0.1 ml sample of each population is transferred to a flask of 9.9 ml of fresh growth medium and thus allowed to keep on evolving (see Figure 1). Every 75 days (every ~500 generations) population samples from each strain are archived in a -80°C freezer, giving researchers a "frozen fossil record" of the populations' evolutionary history (Lenski 2011). At any time, samples from these frozen flasks can be thawed and revived for further analysis of the population at the relevant time point, or even to literally back up and "rerun" a given lineage from an earlier time point.

[3] Except for one neutral genetic marker, discussed below, which distinguishes six lineages from the other six but does not affect fitness.

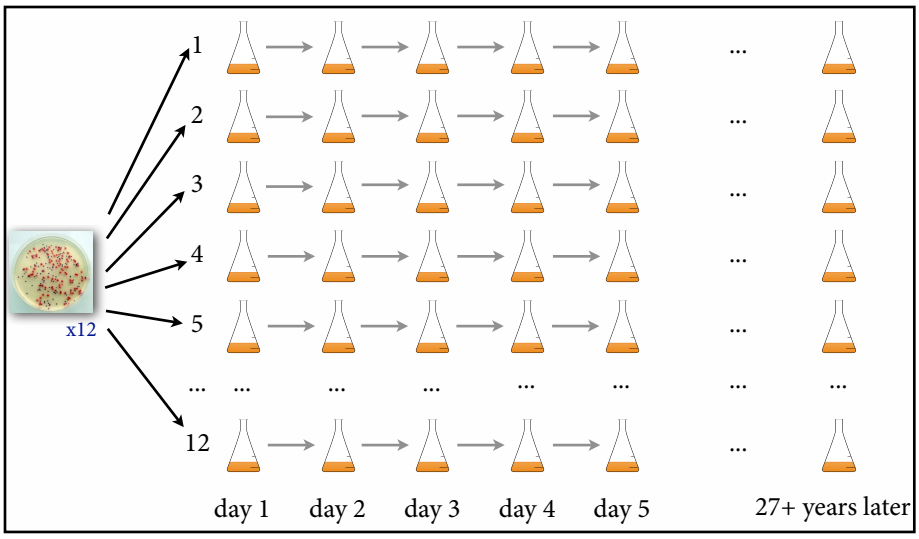

Figure 1: An illustration of the serial transfer protocol used in Lenski's experiment.

The original motivation for Lenski's experiment was to learn about the dynamics of adaptation and diversification by watching evolution in real time. Because the populations started out genetically identical and are evolving in identical environments, any differences in fitness, physiology, or morphology which arise over time are due entirely to new mutations. Since the experiment began in 1988, researchers in Lenski's lab have been keeping the serial transfer protocol going every 24 hours. These populations go through approximately 6.6 generations per day, which means they have a long evolutionary history. In 2010 they passed the 50,000 generation mark, and they are still going today (Lenski 2015). To give some perspective to that number: There is debate about exactly how long our species, *Homo sapiens*, has been around, but estimates point to something on the order of 5,000 to 10,000 generations.[4]

[4] Estimates (depending on method) indicate that the common ancestor of today's humans lived 100,000 to 200,000 years ago (see, e.g., Jorde, Bamshad, & Rogers 1998). This generation calculation is using a conservative estimate of 20 years per generation; genetics studies commonly use 20–25 years (Langergraber et al. 2012).

In addition to the actively evolving populations themselves, researchers have a "frozen fossil record" of the populations' evolutionary history (Lenski 2011). Every 75 days (every ~500 generations), a sample from each well-mixed flask is archived in a –80°C freezer. At any time, these fossil layers can be thawed and revived, by scraping a tiny sample from the frozen tube into a flask of fresh growth medium and growing it overnight at 37°C. This allows researchers to further analyze the revived population at the relevant time point—or even to back up and restart a lineage from an earlier point in its evolutionary history.

One important use of the frozen fossil record is to periodically revive evolved populations and compete them against the original ancestor, which gives a measurement of each population's relative fitness at a given time point in its evolution. Growth rate is used as a proxy for fitness. Here is an overview of how the competitions work. I said earlier that the twelve initial populations were genetically identical; this is true with one exception: The populations have a neutral genetic marker (a marker which does not affect their fitness), which causes them to turn either red or white when grown on tetrazolium arabinose (TA) plates. Six populations have a marker that makes them turn red (the "Ara^-" strains) on TA plates; the other six (the "Ara^+" strains) have a marker that makes them turn white. The evolved strain whose fitness is being measured is mixed in a flask with an equal amount of an ancestral strain with the opposite genetic marker. For example, if the strain whose fitness is being measured is Ara^+, it is mixed with its Ara^- ancestor. A sample from the flask is grown overnight on a TA plate; the number of Ara^+ versus Ara^- colonies can then easily be counted due to their difference in color (Figure 2). After 24 hours have passed, allowing the mixed populations in the flask to compete for resources, samples from the flask are plated again to TA. The growth rate of each population is the natural log

of the ratio of its final plated density (on day 1) to its initial plated density (on day 0). The relative fitness of an evolved strain, then, is a measure of its growth rate relative to that of the ancestral strain during these direct competitions.

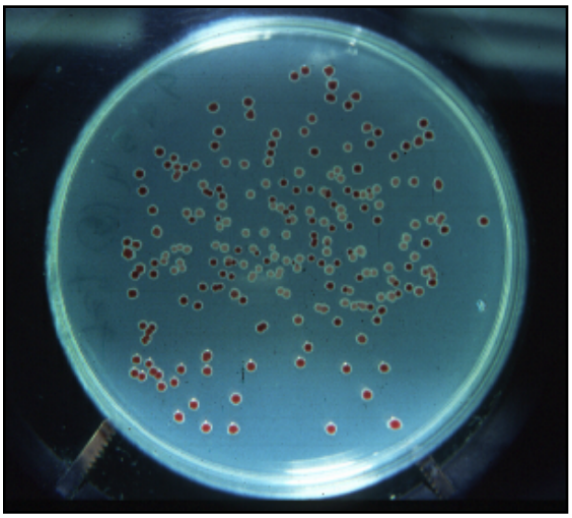

Figure 2: A TA plate showing red and white Ara^{-}/Ara^{+} colonies (photo credit: Paul Sniegowski).

Lenski and colleagues describe their work as "organized around the analogy to an increasingly fantastic exploration of fossil beds" (Lenski & Travisano 1994). In ideal circumstances, a paleontologist can find well-preserved fossil layers which allow them to measure all sorts of features of long-gone lineages and their changes over time. But studying fossil layers in the field leaves many inferences to be made about selection, drift, mutation, and migration, and it leaves open important questions about the populations' environments. The Lenski experiment produces dense and perfectly preserved fossil layers, in the tubes of frozen *E. coli* at 500-generation intervals over tens of thousands of generations. Furthermore, the need to speculate about the populations' evolutionary history and surrounding conditions is eliminated. Researchers have access, at least in principle, to information about which mutations arose and when, whether and when

migration occurred, the exact nature of the environment, and myriad other details inaccessible in fossil records outside the lab. Not only do they have access to this fossil record, but the populations preserved in it can be revived and their evolutionary history can be literally rerun.

Over 60 papers have been published on the Lenski experiment since its beginning. Early papers focused on questions about adaptation and divergence over the first several thousand generations (Lenski et al. 1991; Lenski & Travisano 1994). Later papers have focused on an impressive range of topics in evolution and ecology including the evolved populations' response to novel environments (Travisano et al. 1995), ecological mechanisms promoting coexistence of two populations (Turner, Souza, & Lenski 1996), the relative roles of history, contingency, and selection (Travisano et al. 1995), and three cases which I discuss in more detail in the following subsections: the evolution of high mutation rates (Sniegowski, Gerrish, & Lenski 1997), punctuated morphological evolution (Elena, Cooper, & Lenski 1996), and the evolution of novel traits like citrate utilization (Blount et al. 2012; Blount, Borland, & Lenski 2008). The breadth of what has been learned from this single experimental system is astounding.

Just from this brief discussion of Lenski's experiment, a number of interesting features begin to emerge. The experiment did not begin with a specific hypothesis it was supposed to be confirming or rejecting, but many hypotheses have been tested along the way. Further, the setup seems to incorporate elements of what people tend to consider two very different kinds of inquiry: active manipulation, and passively watching to see what happens while nature runs its course, typically considered the hallmarks of experiment and observation, respectively. The Lenski experiment is happening in an "artificial" laboratory system, but it captures the process of natural selection in its most "natural"

form. These bacterial populations are not undergoing artificial selection or simulating natural selection, they are evolving via natural selection. These features raise a number of questions about the relationship between experiment and theory, the role of experiment in scientific inquiry, and how experimental systems relate to systems in the natural world. I get into these questions in detail in Chapters 4 and 5.

2.1.1. High Mutation Rates

The Lenski populations have been used to study the evolution of high mutation rates over long evolutionary time scales. Natural selection needs variation to act on, and new mutations in organisms' genetic material are the source of that variation. Most new mutations are deleterious, which intuitively makes sense: There are more ways to mess something up at random than there are ways to improve it. One might expect that in well-adapted populations in unchanging environments (that is, in the absence of new selection pressure), genomic mutation rates would stay the same over time or perhaps even decrease. In just three of the twelve Lenski populations, the opposite happened: Their genomic mutation rates increased by two orders of magnitude after 10,000 generations. The explanation for how this happened involves mutator alleles, which raise the genomic mutation rate by inhibiting mechanisms like DNA mismatch repair or proofreading. Sniegowski and colleagues (1997) found that mutator alleles arise spontaneously, by mutation, and then hitchhike to high frequencies[5] in the experimental populations, which is possible because those populations are completely asexual.

[5] Mutator hitchhiking is a phenomenon whereby new mutations arise in a genome which are favored by natural selection and which happen to be linked to mutator alleles. In strictly asexual populations there is no genetic recombination. So if a new mutation arises linked to a mutator allele, and natural selection favors that new mutation, it will spread through the population in subsequent generations and the mutator allele can "hitchhike" along with it, resulting in a population with a higher genomic mutation rate.

From this point about high mutation rates evolving in three of the laboratory populations, the authors make an inference about populations outside of the laboratory. They conclude that high mutation rates might evolve via the same hitchhiking mechanisms in similar clonal populations of cells with high mutation rates in the natural world—in particular, pathogenic *E. coli* and *Salmonella* (Sniegowski, Gerrish, & Lenski 1997, p. 704).

2.1.2. Punctuated Evolution

In this second example, a different sort of inference is made from the same object of study. In a paper called "Punctuated Evolution Caused by Selection of Rare Beneficial Mutations," Lenski and colleagues used results from their evolving *E. coli* populations to make claims about punctuated equilibrium: the notion that evolution occurs in long periods of relative stasis punctuated by short periods of rapid change (Elena, Cooper, & Lenski 1996). Punctuated equilibrium was a hot topic in contemporary discussions of macroevolutionary trends in the fossil record, notably championed by Gould and Eldredge in their discussion of events like the Cambrian explosion, in the context of rejecting views about the ubiquity of phyletic gradualism (Eldredge & Gould 1972; Gould & Eldredge 1977).

Over the first few thousand generations of the Lenski experiment, the twelve populations of E. coli increased in both average fitness and average cell size (Lenski et al. 1991; Lenski & Travisano 1994). In addition to these overall trends, average cell size increased in a step-like pattern. In one population it remained stable for the first 300 generations, increased by over 25% in the following 100 generations, and then remained

stable for another 300 generations before dramatically increasing again (see Elena, Cooper, & Lenski 1996 Figure 1, reproduced as Figure 3 below). This led the researchers to claim that punctuated evolutionary trends could be observed on (relatively) very short time scales, associated with the rise of beneficial new mutations in the population which rapidly sweep to fixation.[6] They conclude the paper with an inference from what happened with the experimental populations of bacteria to a claim about what might be going on in the fossil record, or at least a subset of populations represented in the fossil record, delineated by certain features of their evolutionary history and environmental conditions:

> The experimental population was strictly asexual, which may have increased our ability to resolve punctuated changes. However, any difference between sexual and asexual populations with respect to the dynamic of adaptive evolution breaks down when two conditions are met: (i) standing genetic variation for fitness is exhausted, as will eventually happen in any constant environment, and (ii) beneficial mutations are so rare that they occur as isolated events. To the extent that these conditions are fulfilled in nature, then the selective sweep of beneficial alleles through a population might explain cases of punctuated evolution in the fossil record. (Elena, Cooper, & Lenski 1996, p. 1804)

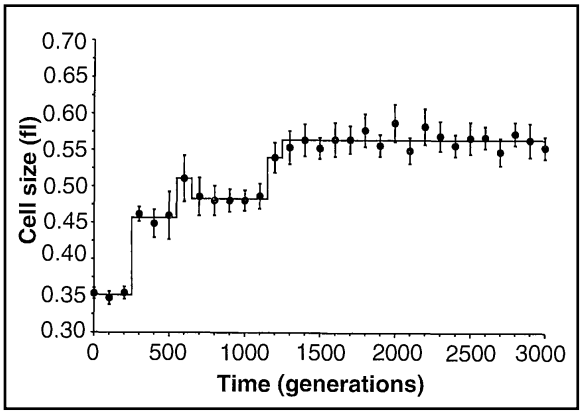

Figure 3: Cell size as a function of time (over 3,000 generations) in one of the 12 populations on the Lenski experiment. From (Elena, Cooper, & Lenski 1996, Figure 1).

[6] The authors are careful to use the term 'punctuated evolution' rather than 'punctuated equilibrium'. Still, the paper is written so as not to avoid being interpreted as saying something about the much more long-term dynamics of punctuated equilibrium involving speciation events. I say more about why this is problematic in Chapter 5.

2.1.3. Citrate Utilization

The glucose-limited growth medium the Lenski populations live in also contains citrate. Citrate is an additional energy source in principle, but *E. coli* cannot use citrate as an energy source in oxic conditions (when oxygen is present). During the first 30,000 generations of the experiment, none of the twelve populations evolved the ability to exploit environmental citrate. As Lenski describes this in a later paper,

> One of the defining features of *E. coli* as a species is that it can't grow on citrate because it's unable to transport citrate into the cell. For 15 years, billions of mutations were tested in every population, but none produced a cell that could exploit this opening. It was as though the bacteria ate dinner and went straight to bed, without realizing a dessert was there waiting for them. (Lenski 2011)

By 31,500 generations, a Cit^+ (citrate utilizing) phenotype had evolved in just one of the 12 strains, strain Ara^{-3}. As the authors of the first paper on this surprising result point out, by 30,000 generations "each population experienced billions of mutations, far more than the number of possible point mutations in the ~4.6-million-bp genome. This ratio implies, to a first approximation, that each population tried every typical one-step mutation many times. It must be difficult, therefore, to evolve the Cit^+ phenotype, despite the ecological opportunity" (Blount, Borland, & Lenski 2008, p. 7900).

Blount and colleagues tested whether the Cit^+ phenotype arose in strain Ara^{-3} because of an unusual rare mutation, or because of historical contingency and the particular evolutionary trajectory that strain had taken. They did this by "replaying" Ara^{-3} many times from time points prior to the 31,500 generation mark, using the same sort of defrost-and-revive protocol described above. The rare-mutation hypothesis would predict that Cit^+ phenotypes evolve at the same very low rate from any previous time point; the historical-contingency hypothesis would predict that the mutation rate to Cit^+ phenotypes increases after some series of potentiating mutations occur (that is, it increases as the

evolutionary time point approaches 31,500 generations). They replayed and tested hundreds of replicates of strain Ara^{-3} from earlier time points, and the results supported the historical contingency hypothesis: some genetic background had arisen earlier in that strain's history which paved the way for the Cit^{+} phenotype.

Concurrently with the Cit^{+} phenotype arising, there was also significant population expansion in Ara^{-3}, measured by optical density readings of the strain before and after the 31,500-generation time point. Interestingly though, the Cit^{+} phenotype did not fix in that population; Cit^{-} (non-citrate utilizing) phenotypes persisted at a frequency of around 1%.[7] The authors suggest that the Cit^{-} cells persist as glucose specialists, given their higher growth rate and shorter lag phase compared to Cit^{+} cells when utilizing glucose (Blount, Borland, & Lenski 2008). Frequency-dependent selection here appears to maintain ecological diversity in the population. The paper suggests future directions for research on the evolution of this novel phenotype including studying whether differentiation into Cit^{+}/Cit^{-} phenotypes might be a first stage in sympatric speciation.

2.1.4. Laboratory Natural Selection?

Before moving on, it is worth addressing and setting aside a question—are these laboratory populations really undergoing natural selection? Two related ideas might motivate this question. The first holds that struggle for existence is a necessary condition for natural selection (as Darden has argued, in a personal communication, but see also Darden & Cain 1989) and questions whether there is genuine struggle for existence in the Lenski system. Second, one might wonder whether natural selection is really acting when

[7] This stable coexistence was confirmed in subsequent replays of mixed Cit^{+}/Cit^{-} populations; different initial concentrations of the two all eventually converged to the same equilibrium frequency of around 1% Cit^{-}, 99% Cit^{+}.

researchers are imposing the conditions, as part of the serial transfer protocol, for which organisms carry on to each subsequent generation.

In response to the first point: Questioning whether or not natural selection is acting here, for this reason, implies subscribing to the view that there can be struggle for existence only when that struggle is between organisms and a changing, challenging abiotic environment. On this line of thought, the abiotic environment (the bacterial growth medium) in Lenski's experiment, kept constant throughout the experiment, presents the bacteria with no opportunity for struggle for existence and thus removes a necessary condition for evolution via natural selection. This is wrong for two reasons. First, while the recipe for the nutrient broth is always the same, the abiotic environment in the flask changes over the course of every 24 hours as the populations deplete environmental glucose. Second, and more importantly, the environment relevant to thinking about the struggle for existence includes not only abiotic factors but other organisms as well. The citrate utilization case (Section 2.1.3) is a perfect illustration of this: Citrate was part of the abiotic environment all along, but the sudden arrival of conspecifics capable of using it to their advantage changes the selective environment for the population as a whole. Darwin himself understood struggle for existence the way I am here, to include other organisms as well as the abiotic environment:

> I use the term Struggle for Existence in a large and metaphorical sense, including dependence of one being on another, and including (which is more important) not only the life of the individual, but success in leaving progeny. Two canine animals in a time of dearth, may be truly said to struggle with each other which shall get food and live. But a plant on the edge of a desert is said to struggle for life against the drought, though more properly it should be said to be dependent on the moisture. A plant which annually produces a thousand seeds, of which on an average only one comes to maturity, may be more truly said to struggle with the plants of the same and other kinds which already clothe the ground… In these several senses, which pass into each other, I use for convenience sake the general term of struggle for existence. (Darwin 2009, pp. 62–3).

In response to the second point, about it being up to researchers which individuals will carry on to the next generation: Population samples are taken from well-mixed flasks and transferred to a new flask each day. Researchers are not choosing which individuals from the population will make it into the next flask based on any features those individuals possess, such as their phenotype or their location in the flask. The flask is constantly shaken over the course of the experiment, so the sample that carries on to the new flask each day is random and representative of the population. A bottleneck occurs each time a sample is transferred to a flask of fresh growth medium, but nobody has any say in which organisms make it through the bottleneck.

The selection process in the Lenski experiment is really natural selection. That is a crucial feature of what makes this and other LNS experiments such powerful tools for studying evolution in action. There might be all sorts of other reasons, in a given case, why inferences from what happens in the laboratory to what happens in nature might call for validation. As I will discuss in Chapter 5, some such inferences are more questionable than others due to differences in our reasons to trust that the laboratory populations are relevantly similar to the target populations in question. But "It's not really natural selection" is not one of these reasons.

2.1.5. The Lenski Experiment?

Another question that might come up about the Lenski experiment is: What is "the experiment?" The 27+ years (tens of thousands of generations) of propagation of replicate *E. coli* populations are commonly referred to as "the Lenski experiment." However, many different treatments have been applied to those populations over the years, discussed in

the above-mentioned 60+ publications from Lenski's group based on this work.[8] It might also make sense to think of the particular treatments and protocols contained in each of those papers as an experiment or set of experiments in itself (the researchers themselves sometimes refer to new treatments concurrent with the ongoing evolution experiment, like varying the growth medium or temperature, as separate evolution experiments (Lenski 2011)). So, there might be an argument for thinking of those studies as the experiments, and the 26+ years of propagation as a research tool rather than an experiment in itself. I suggest not getting hung up on worrying about the ontology of particular experiments at this level of detail. Nothing rests on this for the purposes of this dissertation. It makes more sense to talk about cases of experimental inquiry—Lenski's work certainly counts—rather than looking for clean lines to draw around one experiment and another within the context of a given scientist's or laboratory's research.

2.2. LNS Experiments Beyond Lenski: History and Current Directions

The Lenski experiment is especially impressive and deservedly famous. But LNS experiments actually date back to the 1880s, when William Henry Dallinger reported on a long-term evolution experiment in his Presidential Address to the Royal Microscopical Society (Dallinger 1887; see also discussion in Huey & Rosenzweig 2009). Dallinger's goal was to discover "whether it was possible by change of environment, in minute life-forms, whose life-cycle was relatively soon completed, to superinduce changes of an adaptive character, if the observations extended over a sufficiently long period" (1887, p. 191). His

[8] See http://myxo.css.msu.edu/PublicationSearchResults.php?group=aad.

subjects were "the lowest forms of the infusoria,"[9] and his chosen environmental manipulation was slowly increasing the temperature. He constructed an apparatus designed to maintain a constant temperature: a large water-filled copper vessel insulated with felt, with space for three glass inserts containing "putrefactive fluids" and the microorganisms themselves (Figure 4). This was attached to a mercury thermometer and gas flame rigged to automatically regulate and stabilize the temperature in the glass vessels to within 1/4 of a degree Fahrenheit. Dallinger began slowly turning the temperature up—from 60° to 70°F (15.6°–21.1°C) over the first four months of the experiment—and observed the organisms in the vessels through a microscope attached to the apparatus, noting changes in their morphology and behavior. He gives a detailed account of their slow adaptation and stabilization at various temperatures as he continued to turn up the temperature after those first four months, describing them at 93°F (33.9°C) as about "to surrender to torpor and death," after which point some organisms seemed to stabilize. With painstaking diligence he describes slowly turning the temperature up, with long pauses in between, finally reaching 158°F (70°C). He concludes with a note on the remarkable adaptation of his populations at later time points: "If the adapted organisms at 158° F. were taken from that temperature and placed in an eminently nutritious and suitable nutritive fluid at 60° they died. While, of course, if forms of the same kind exactly, living and flourishing at 60°, were placed in a nutritive sterilized fluid at even 150° they were finally destroyed" (p. 199).

[9] 'Infusoria' is an obsolete term for a class of aquatic microorganisms, including primarily the organisms we now classify as protists. Dallinger focused on three particular species of flagellated protists.

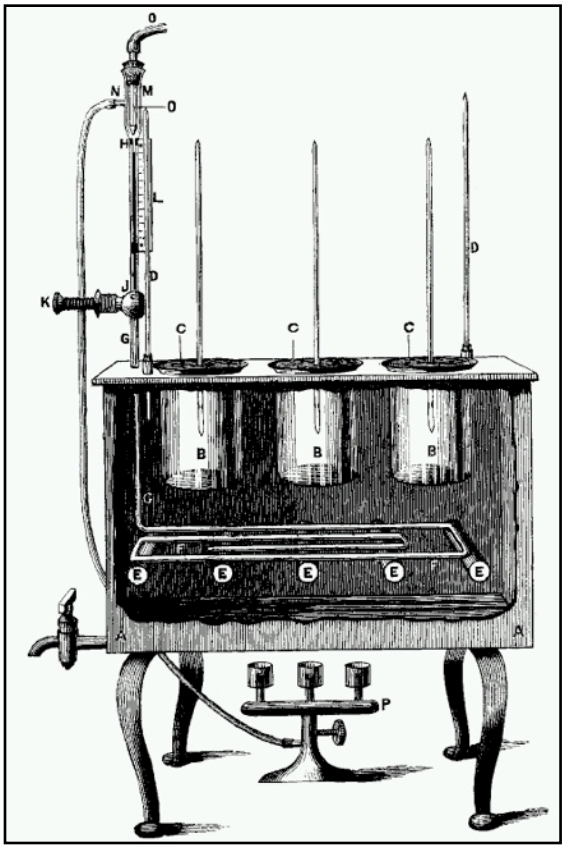

Figure 4: Dallinger's drawing of the apparatus used in his evolution experiment (from Dallinger 1887).

Dallinger's experiment took seven years. It was brought to an abrupt halt only when "an accident, which no foresight could have guarded against, happened to the apparatus employed" (p. 190).[10] This stands out as a remarkably early exemplification of the key features of LNS experiments discussed above: long-term study of populations in the laboratory evolving via natural selection in a controlled environment.

LNS experiments began using bacteria in the 1950s, when Novick and Szilard (1950) used a chemostat to limit resources in populations of *E. coli*, controlling their

[10] A side note of interest: Dallinger corresponded with Darwin in 1878 about his preliminary results; he quotes a letter in which Darwin wrote to him: "I did not know that you were attending to the mutation of the lower organisms under changed conditions of life; and your results, I have no doubt, will be extremely curious and valuable. The fact which you mention about their being adapted to certain temperatures, but becoming gradually accustomed to much higher ones, is very remarkable. It explains the existence of algae in hot springs. How extremely interesting an examination under high powers on the spot, of the mud of such springs would be" (Dallinger 1887, pp. 191–192).

growth rate, and observed their rates of spontaneous mutation to resistance to bacteriophages T_4 and T_5. Another notable early LNS experiment, published a year later, is described in a paper by Atwood and colleagues propagating *E. coli* in Erlenmeyer flasks at 37° C and transferring samples from each flask to a new one every 12 hours (1951). *E. coli* are popular organisms of choice for the reasons cited above: their short generation times, ease of manipulability, ability to freeze and revive, and the wealth of existing knowledge about their genomes and other traits. But LNS experiments use all sorts of other organisms as well, including viruses (Bono et al. 2013; Forde & Jessup 2009), fungi (Gifford, de Visser, & Wahl 2013), yeast (Gerstein 2013; Ratcliff et al. 2012; Zeyl 2000), green algae (Bell 2013; Colegrave et al. 2002), nematodes (Matsuba et al. 2013), insects (Simoes, Santos, & Matos 2009; Zera & Harshman 2009), and mice (Barnett & Dickson 1984).

Since the early examples discussed above and the beginning of Lenski's experiment, LNS experiments have made progress in a variety of issues in evolutionary biology and ecology. I do not have space for a thorough overview (but see Bataillon, Joyce, & Sniegowski 2013; Garland & Rose 2009 and papers therein), but will mention just a few more brief examples here to illustrate the breadth of topics covered in the field beyond Lenski's work.

Major transitions in evolution have been a hot topic among philosophers of biology and biologists (e.g., Calcott & Sterelny 2011; De Monte & Rainey 2014; Maynard Smith & Szathmary 1997), including notably the evolution of multicellularity. Recent experiments have contributed to our understanding of multicellularity by trying to evolve it de novo in the lab. These include Michael Travisano and colleagues' experiments evolving multicellular yeast, in which they expose yeast to regular selection pressure for

sinking to the bottom of test tubes, and watch it begin to form "snowflake"-like clusters which appear to show some division of labor, characteristic of entities at the transition from unicellular to multicellular life (Ratcliff et al. 2012). Rainey and colleagues have examined the transition in experimental populations of the bacterium *Pseudomonas fluorescens*, in a spatially structured heterogenous environment, from living as individuals to living in cooperating groups, a first step in the evolutionary transition to multicellularity (Rainey & Rainey 2003). This work on experimental evolution of multicellularity ties to questions in the philosophy of biology about levels of selection and individuality, for example, at which point do these multicellular clusters become individuals or objects of selection in themselves, and what are the implications?

Recent experimental microbial evolution studies have also investigated the evolution of sex. Sexual reproduction is more costly than asexual reproduction. As Maynard Smith famously put it, there is a "twofold cost of sex:" Organisms that have sex contribute only half of their genetic material to their offspring (versus asexual organisms which contribute all of it), and all else being equal an asexual population grows twice as quickly as a sexual one (Maynard Smith 1978). Recent work in experimental evolution has investigated the tradeoffs between asexual and sexual reproduction. Raynes and colleagues investigated the effects of recombination on the frequency of mutator alleles in experimentally evolved populations of yeast, comparing asexual to sexual populations (Raynes, Gazzara, & Sniegowski 2011). Turner and colleagues review recent LNS studies of viruses, yeast, *E. coli*, and *Chlamydomonas* (green algae) establishing that sexual (recombining) populations adapt faster than their asexual counterparts (Turner, McBridge, & Zeyl 2009).

LNS experiments have also been used to study the mechanisms maintaining biodiversity. Kerr and colleagues have used laboratory microbial populations to look at the stable coexistence of different strains of *E. coli* in different spatial configurations. They set up a rock-paper-scissors dynamic involving three strains of *E. coli*: a colicin-producing strain (C) produces a toxin which kills a colicin-sensitive (S) strain; the sensitive strain however outgrows a third colicin-resistant strain (R), which in turn outgrows the colicin-producing strain. Thus there is a rock paper scissors dynamic: C beats S; S beats R; R beats C. They found that the three stains stably coexist when their interactions are spatially localized, but not when they disperse and interact over relatively large spatial scales (Kerr et al. 2002).

LNS experiments are making progress in a host of other research areas including studies of the interaction between evolution and ecology (Jessup et al. 2004), our understanding of cancer as an evolutionary phenomenon (Sprouffske et al. 2012), genetic interactions and their influence on evolution, the evolution of sociality, and how environmental variation affects evolution (see Bataillon, Joyce, & Sniegowski 2013; Garland & Rose 2009 and papers therein).

2.3. Tapes of Life, Darwin's Mistake, and Other Challenges

Darwin believed that evolution happened very slowly, and could not be directly observed. Of evolution in action, he wrote: "We see nothing of these slow changes in progress, until the hand of time has marked the long lapse of ages, and then so imperfect is our view into long past geological ages, that we only see that the forms of life are now different from what they formerly were" (Darwin 2009, p. 84). As Rose and Garland argue in their

provocatively titled paper "Darwin's Other Mistake,"[11] experimental evolution shows that this is wrong (see also Bataillon, Joyce, & Sniegowski 2013). It allows us to watch evolution as it happens, and observe remarkable changes in populations with generation times shorter than our own.

Another idea which LNS experiments have challenged regards our ability to "replay the tape of life." As part of his argument for evolutionary biologists to take historical contingency more seriously, Gould proposed a thought experiment in which we could rewind the history of life on Earth to some point in the deep past, erase what had happened from that point on, and replay it to see how things evolved. Gould said that if we were to do this, we could expect that "any replay of the tape would lead evolution down a pathway radically different from the road actually taken" (1989, p. 51). This is a powerful thought experiment, and Gould presented it as only a thought experiment, saying that we could never actually replay the tape of life. Lenski's long-term evolution experiment proves him wrong: not his claim that evolution would differ were we to replay the tape, but his claim that we could never perform this experiment on real living systems by backing them up and "replaying" them from some earlier point in their evolutionary history, from the same starting conditions. Beatty (2006) and Desjardins (2011, note 3) have noted in previous discussions of the Lenski experiment that "playing" the twelve different lineages from the exact same genetic starting point in identical environments gives us an approximation of what Gould had in mind with replaying the tape. In other words, we can think of the twelve lineages as twelve "runs" of the same course, since they all began with the same ancestor. But there is also a further sense in which long-term microbial evolution experiments like Lenski's allow us to replay the tape of life: As

[11] The first mistake alluded to is Darwin's set of beliefs about the mechanism of heredity and blended inheritance.

discussed above, the experimental populations can be frozen, defrosted, and "rerun" from any point in time. In other words, not only does the set of twelve lineages represent twelve replays of the same tape, the individual lineages themselves can be replayed, just by defrosting them and regrowing them. This latter aspect of the experiment represents an even more straightforward way in which experimental evolution allows us to replay the tape of life. Gould had in mind replays of the tape of life on a much vaster evolutionary time scale, spanning speciation events and hundreds of thousands of generations or more. The evolutionary time scale of the Lenski experiment is (so far) orders of magnitude smaller. In any case, the long-term evolution experiment offers an extremely powerful tool for investigating questions about the relative roles of history, chance, and adaptation in evolutionary processes (Futuyma & Bennett 2009; Travisano et al. 1995).

The Lenski experiment is by far the most well-known LNS experiment. It is the only one philosophers have talked about in detail. Beyond Beatty and Desjardins' discussions, philosophers had not had much to say about this research area.[12] In the rest of this dissertation I return frequently to examples of LNS experiments, as a case study for challenging ways philosophers and scientists have talked about methodological boundaries in science.

What is special about LNS experiments, that they seem to straddle these different classic divisions in scientific methodology? We generally think of the ways to do biology as doing theory, studying laboratory-bred populations, or studying populations in nature. (This is not meant to be an exhaustive list, but the main research methodologies that come

[12] There are a few exceptions; for example Love and Travisano (2013) discuss examples of experimental microbial evolution in their recent paper on microbes as model organisms in developmental biology, and Abrams (2012) mentions the Lenski experiment in a broader discussion of different ways to think about fitness. There have also been sessions at recent meetings of the Philosophy of Science Association and The International Society for the History, Philosophy, and Social Studies of Biology focused on experimental evolution.

to mind.) These all have their respective advantages, and criticisms of the others: theoretical biology lets you control whatever variables you want in your models but it is not realistic enough; laboratory experiments allow for controlled interventions but the systems of study are too artificial; field work lets you genuinely study nature in action but lacks control. LNS experiments are a special case: They take place in the laboratory, allowing for controlled intervention, and the populations of study are "artificial" in the sense that they are strains adapted (or adapting) to the laboratory environment. But they are doing something that natural populations typically do and laboratory experimental populations typically do not: evolving via natural selection. And we can interact with them in this particular and very powerful way in which we could previously only interact with computer simulations: We can back them up and rerun them. In addition to making LNS experiments a powerful and exciting research area for biologists, these features give philosophers interested in biology, scientific methodology, and related epistemological questions plenty to think about. I return to some of these themes later in the dissertation.

3. Hypothesis-Testing and Exploratory Experiments

As discussed in Chapter 1, in the last few decades philosophers and historians of science have pushed to focus more attention on experiments and their role in scientific inquiry (A. Franklin 1990; Galison 1987; Hacking 1983; Radder 2003; Weber 2004). Recent literature on exploratory experiments has addressed ways in which the classic picture of hypothesis-testing does not capture everything that experimenters are up to (Burian 2007; Elliott 2007; L. Franklin 2005; O'Malley 2007; Steinle 1997; Waters 2007). In this chapter, I argue that we need a better account of the whole space of experimental inquiry; hypothesis-testing and open-ended exploration are not the entire picture. Waters (2007) and Elliott (2007) have previously made the point that there is such a space, but much work remains to be done to clarify it. As a starting point for that larger project, I propose a more nuanced way to think about the relationship between experiment and theory.[13] I will frame this discussion within the context of examples of LNS experiments. This is a good starting point because these examples put pressure on the idea that cases of experimental research can be neatly classified as one or the other (hypothesis-testing or exploratory).

In Chapter 1 I talked about some classic examples of experiments, Redi's experiments on insect generation and the Meselsohn-Stahl experiment. From the discussion in Chapter 2, it should already be clear that LNS experiments are different from these canonical cases in interesting ways. The Lenski experiment did not begin with the aim of confirming or rejecting a particular hypothesis, but a number of questions have

[13] Let me be clear from the outset what I mean by 'theory'. I intend this broadly to encompass both syntactic and semantic understandings of theory: Theory can comprise a lawlike series of axioms, or a collection of models. I like the way MacLeod and Nersessian characterize theory: "[A] broad eclectic understanding of theory as a reservoir of laws, canonical theoretical models, principles of representation… and ontological posits about the composition of phenomena." (2013, p. 539).

been answered, and hypotheses tested, along the way. Furthermore, the experimental setup incorporates elements of what we usually consider two very different kinds of inquiry: active intervention and manipulation, versus passively watching to see what happens while nature runs its course, typically considered the hallmarks of experiment and observation, respectively. The Lenski experiment takes place in an "artificial" laboratory system, but it was motivated by the desire to capture the process of natural selection in its "natural" form; these bacterial populations are not undergoing artificial selection or simulating natural selection, they are really evolving via natural selection. These sorts of features raise questions about the nature of experimental inquiry and the relationship between experiment and theory. In the rest of this chapter, I lay out the two main views in the literature on the nature of experimental inquiry and the relationship between experiment and theory, and discuss why neither of them alone adequately characterizes what is going on in Lenski's experiment. This points to the need for a more nuanced understanding of possible experiment–theory relationships, as more than just a binary set of options. The latter point is not novel (see discussion below of Brandon's, Waters', and Elliott's contributions on this front). My contribution in this chapter is to make progress on clarifying the conceptual space of experimental inquiry by developing an account of how we should think about a key set of dimensions of that space: the relationship between experiments and theories.

Sections 3.1 and 3.2 discuss the views of experimental inquiry as hypothesis-testing and exploration, respectively, and discuss why neither of these alone captures what is going on in Lenski's experiment. Section 3.3 gives an overview of previous attempts to map a conceptual space of experimental inquiry, including the mostly ignored middle ground between pure exploration and classic hypothesis-testing. In Section 3.4 I outline a

proposal for how to better approach mapping a key aspect of the space including this middle ground, focusing on the roles that theories play in experiments.

3.1. Hypothesis Testing: The Classic View of Experiments

Philosophers of science have traditionally focused on theories and models as the fundamental subjects of analyses of scientific inquiry. Experiments came into play as means for linking theories and models to the world. On the classic view, theories are the focal starting point, both conceptually and in practice: Scientists begin with their theory in hand, formulate a particular hypothesis based on that theory, and then design a particular experiment whose aim is to test that hypothesis. As mentioned in Chapter 1, Hempel's writings on experiments are a prime example of this hypothesis-testing view (Hempel 1966). According to Hempel, theories are central in scientific inquiry. Theories give rise to hypotheses with the form "under conditions *C*, events of type *E* will occur," and the point of experiments is to bring about conditions *C* and check whether or not events of type *E* do, in fact, occur. Thus, the business of experiments is hypothesis-testing, and the best experiments are what Hempel calls "crucial tests:" experiments designed to settle the conflict between two rival hypotheses about some phenomenon in the world which have thus far stood up equally well, until the crucial-test designer identifies a situation in which they predict different experimental outcomes. Hansen has a similar view of experiments as hypothesis-testing, when he writes that good experiments test "single, tersely-expressed hypotheses" (1958, p. 67). In a similar vein, Popper wrote in his *Logic of Scientific Discovery* that "The theoretician puts definite questions to the experimenter, and the latter by his experiments tries to elicit a decisive answer to these

questions, and to no others… Theory dominates the experimental work from its initial planning up to the finishing touches in the laboratory" (Popper 1934, p. 107).

This classic view of experiments as hypothesis-testing was developed with examples from physics in mind, and it applies well to many classic and contemporary examples from physics. Part of the pushback against this traditional view has stemmed from the objection that it does not apply as well to other sciences, like the life sciences. But it is worth noting that this does not mean that the hypothesis-testing view never captures what is going on in biology. Plenty of biology experiments follow the hypothesis-testing model. The Meselsohn-Stahl experiment was one example; here is another from my work in the Sniegowski laboratory at Penn. This case involves questioning the existence of general antimutators: alleles that lower the occurrence of mutations across an organism's genome by improving the accuracy of mechanisms like nucleotide insertion, proofreading, or DNA mismatch repair. Antimutators are known to exist in some organisms like the bacteriophage *T4*, but they are specific, not general. That means that they do not reduce the occurrence of mutations across the entire genome, but instead reduce mutations only in particular sites on the genome, or of particular types (for example, transitions but not transversions).[14] In his paper "General Antimutators are Improbable," Drake (1993) argues against the existence in principle of non-specific antimutators. For both structural and evolutionary reasons, he doubts that general antimutators that lower an organism's entire genomic mutation rate could exist.

Some studies claim to have identified general antimutators, with reported strengths ranging from 3- to 50-fold decreases in per-base-pair mutation rate (Fijalkowska

[14] The four DNA bases—adenine (A), cytosine (C), guanine (G) and thymine (T)—are classified by their chemical ring structures as one-ringed pyrimidines (C and T) or two-ringed purines (A and G). A transition is a mutation where one purine is substituted for another, or one pyrimidine is substituted for another. A transversion is a mutation where a purine is substituted for a pyrimidine or vice versa.

& Schaaper 1993; Quinones & Piechocki 1985). But none of them have been tested for effects on mutation rate outside of the genetic background in which they were isolated. In other words, researchers claim to have found an allele which acts as a general antimutator, but they demonstrated it in action in only one context, the genome of the particular organism in which they identified it. This raises questions about the background specificity of alleged antimutators. They might have a general antimutator effect only in one particular genetic background. A particularly well-characterized example is *dnaE911*, which has been shown to have antimutator effects in *E. coli* K12. The claim being tested is that *dnaE911* is not a genuine antimutator in the broad sense that organisms bearing it will have a lower genomic mutation rate, but that it just happens to have this effect in the particular strain of *E. coli* that Schaaper and colleagues were looking at. To test this, we inserted *dnaE911* into several other strains of *E. coli* and checked whether or not it affected their genomic mutation rates (Gentile, Shaver, & Parke, in preparation). It had no effect on the genomic mutation rate of the other strains of *E. coli* into which we successfully incorporated it, indicating that care should be taken to check for background-specificity in identifying general antimutators.

This is one example from biology that fits the picture of experiment as hypothesis-testing. But this picture does not capture what is going on in all experimental inquiry in biology. The Lenski experiment is a case in point. Unlike the antimutator experiment, the Lenski experiment was not designed to test a particular theoretical prediction. It was informed by the theoretical backgrounds of evolutionary theory and population genetics. But it was motivated by the insight that evolving bacterial populations in the lab in real time, for a long time, would offer an unprecedented and powerful way to learn about the long-term dynamics of adaptation and diversification. As discussed in Chapter 2, while

the experimental results initially centered around the relationship between fitness and cell size, as the experiment went on they incorporated background from, and asked and answered questions about, long-term evolutionary dynamics drawing on areas of theoretical background which were not explicitly drawn on at the experiment's outset, including the dynamics of punctuated equilibrium, beneficial mutations, evolution of mutation rates, ecology, and many others.

Strict hypothesis-testing is not the whole picture; experimental inquiry can be more open-ended. The last couple of decades have seen attempts to give a richer picture of experimental inquiry, and account for the fact that not all experiments are in the business of straightforward hypothesis-testing.

3.2. Exploratory Experiments: A Response to the Classic View

> Isaac Newton argued that "hypotheses… have no place in experimental philosophy," a view echoed by mathematician Roger Cotes: "Those who assume hypotheses as first principles of their speculations… may indeed form an ingenious romance, but a romance it will still be" (in I. B. Cohen Introduction to Newton's Principia; iUniverse, 1999). (Glass 2014)

A recent cluster of literature has aimed to show that previous discussion in philosophy of science has portrayed the scope of experimental inquiry too narrowly. By focusing on theories as central to scientific inquiry, and on experiments purely as means to test the hypotheses theory generates, philosophers have not paid enough attention to experiment in general, and in particular have neglected the existence of an important kind of experiment, exploratory experiments. People have said different things about what exploratory experiments are up to; there is no consensus on how exactly to define them. In this section I overview the literature on exploratory experiments, which has focused primarily on experiments in the life sciences and biotechnology. I then go on to argue

that, while some aspects of the exploratory experiment account capture what is going on in LNS experiments, the latter look quite different from the paradigm examples of exploratory experiment in important ways. More work is needed to fully capture the range of kinds of experimental inquiry that take place in biology, and hence in science in general.

The recent literature I will discuss is the first to explore the idea of exploratory experiments in detail, but these authors are not the first to point out that experiments do more than test hypothesis. Hempel briefly gestures to this in *Philosophy of Natural Science*: "Experimentation… is used in science not only as a method of test, but also as a method of discovery… where no specific hypotheses have as yet been proposed, a scientist may start with a rough guess and may use experimentation as a guide to a more definite hypothesis" (1966, p. 21). Another early objection to the view of experiments as (purely) hypothesis-testing can be found in Hacking's book *Representing and Intervening*, where he discusses the relationship between theory and experiment. Assessing a representative statement of the deductive method in science made by Justus von Liebig in 1863, he writes:

> There is however a strong version of Liebig's statement. It says that your experiment is significant only if you are testing a theory about the phenomena under scrutiny. Only if, for example, Davy had the view that the taper would go out (or that it would flare) is his experiment worth anything. I believe this to be simply false. One can conduct an experiment simply out of curiosity to see what will happen. Naturally many of our experiments are made with more specific conjectures in mind… [but] must there be a conjecture under test in order for an experiment to make sense? I think not. The physicist George Darwin used to say that every once in a while one should do a completely crazy experiment, like blowing the trumpet to the tulips every morning for a month. Probably nothing will happen, but if something did happen, that would be a stupendous discovery. (Hacking 1983, p. 154)

As Hacking goes on to point out, to say that experiment can precede theory is not to say that experiment could exist completely independent of theory. However, it must be

acknowledged that "much truly fundamental research precedes any relevant theory whatsoever" (Hacking 1983). Sometimes the experiment is done and the theory brought to bear or developed later. Thus, even these "simply out of curiosity" experiments are not isolated from theory. But there is a big difference between having this kind of relationship with theory, and being strongly theory-driven in the way Hempel, Hansen, and Popper were talking about. Hacking's is an early version of discussion of what are now called exploratory experiments (A. Franklin 1990; see also discussion in Galison 1987).

Steinle (1997) is usually credited with being the first to explicitly discuss exploratory experiments. He defines them in terms of a particular combination of experimental methodology and absence of relevant theoretical background. In particular, he says that exploratory experiments involve varying many parameters at once with the goal of finding empirical rules, under conditions where a theoretical and conceptual framework is unavailable or unreliable. Subsequent definitions of exploratory experiment have similarly focused on their relationship with theory. Waters (2007, p. 5) defines them as experiments that aim "to generate significant findings about phenomena without appealing to a theory about these phenomena for the purpose of focusing experimental attention on a limited range of possible findings." Franklin-Hall (L. Franklin 2005) defines them as experiments which are not guided at all by hypotheses or by theory.

Thus, the consensus negative definition of exploratory experiments is "experiments that are not engaged in hypothesis-testing." There is no consensus on a positive definition. Most of the literature focuses on particular case studies, highlights ways that these cases do not fit the hypothesis-testing view, and characterizes them instead as paradigm examples of exploratory experiment. A general sense of the common features of exploratory experimentation can be drawn from these cases and surrounding

discussion of their features. These paradigm cases of exploratory experiments all come from new, rapidly-developing research areas in the life sciences, including experiments on microRNAs (Burian 2007), nanotoxicology (Elliott 2007), fMRI (Bateman 2012; L. Franklin 2005), metagenomics (O'Malley 2007), and high-throughput systems biology such as DNA microarray research (L. Franklin 2005).

Three features unite these paradigm examples:

1. being **theory-informed** rather than theory-driven*;*
2. working in the context of **underdeveloped areas of background theory**; and
3. using so-called "**wide instrumentation**."

This is not an exhaustive list of the features associated with exploratory experiment, but I believe this set of three features represents those unanimously endorsed as characteristic of paradigm cases of exploratory experiment. I will say a bit about each of these features in turn.

Everyone talking about exploratory experiments says that they are not theory-driven. Waters (2007) introduced the term 'theory-informed' to capture what theory does to exploratory experiments rather than drive them. Experimentation does not typically happen in complete isolation from theory, so to say that experiment is not theory-driven is not to say that theory plays no role. Theories are in the background motivating and guiding the design of paradigm exploratory experiments, they are just not being used to generate specific questions whose specific answers are the experiment's primary ends. Franklin-Hall (L. Franklin 2005) makes a related point about the role of theory in exploratory experiment by distinguishing *theoretical background* from *local theory*, where theoretical background is the canon of theory that broadly informs and guides an

experiment, while local theory involves specific predictions about the particular entities and processes at play in a particular experimental system. She is not always clear about the difference between theoretical background and local theory, but this helps: "Background theories, among other things, direct inquirers to the kinds of properties that could possibly have a causal role in their local investigations, even if they do not posit particular causal relationships;" the latter is what local theories do (L. Franklin 2005, p. 893). So, saying exploratory experiment is theory-informed rather than theory-driven means that theory still plays an important role, but not the role of driving the experimenter to test a specific prediction. The distinction between theoretical background and local theory helps highlight this relationship: All experiments engage with theoretical background, but theory-driven experiments test hypotheses that fall out of the relevant local theory, while exploratory experiments do not. Their aims with respect to theory are different, and involve things like characterizing new phenomena or generating new theory.

The second feature common to the paradigm cases of exploratory experiments is that they occur in fields with theoretical backgrounds that are new, underdeveloped, rapidly expanding, or some combination of the three. Steinle identifies exploratory experiment as "typically tak[ing] place in those periods of scientific development in which —for whatever reasons—no well-formed theory or even no conceptual framework is available or regarded as reliable" (1997 p.S70). O'Malley (2007) describes how the development of experimental work in metagenomics was motivated by theoretical and practical difficulties with identifying and classifying individual microbes.[15] Elliott's example of nanotoxicology (2007) similarly involves experimental work aimed at

[15] Metagenomics studies analyze the collective genomic material of the microbes found in an environmental sample, allowing researchers to collect large amounts of genomic data and then see what genetic material is there and what they can learn from it, without worrying about which particular organism(s) it belongs to.

generating lots of information to address a gap in theory, in this case, knowledge of the health effects of the new and rapidly expanding body of nanotechnology products. These and the other paradigm cases share the feature of experimental inquiry taking place in the absence of particular theoretical background on the experiment's particular subject matter, and with the explicit goal of eventually helping to flesh out that theoretical background. This characteristic is what motivates Burian to equate exploratory experimentation with "discovery science" (2007, p. 12).

The third common feature of exploratory experiment is using what Franklin-Hall (L. Franklin 2005) has termed "wide instrumentation." This is basically another term for high-throughput instrumentation, techniques that allow researchers to rapidly measure many different features of an object of study in parallel. This goes hand-in-hand with the view of exploratory experiments as productive activities, producing as much data as possible from as many angles as possible to bolster theoretical background where it is lacking. Franklin's discussion of research using DNA microarrays exemplifies wide instrumentation's central role in exploratory experiment. The goal in these experiments was to survey and catalog all of the genes in a yeast genome that had certain cell-cycle-regulating features, for the purposes of helping scientists learn more about transcriptional control in the cell cycle. They were not asking specific questions about a specific aspect of transcriptional control, but rather trying to gather as much information on it as they could from many angles at once. See (Burian 2007; L. Franklin 2005) for further discussion of the connection between exploratory experiment and wide instrumentation.

To summarize what has been said so far in this subsection: Exploratory experiment has been defined negatively as experimental inquiry that is not hypothesis-testing. Positive accounts focus on describing paradigm examples of exploratory

experiment and highlighting their central features. The three features common to these paradigm cases are being theory-informed but not theory-driven, lacking well-established theoretical background, and employing wide instrumentation. It is important to note that these are not meant to be necessary or sufficient conditions for categorization as an exploratory experiment. People say many things about exploratory experiments in the literature and do not agree on a single positive definition; the focus is on examples and their common themes.

To the extent that the three-feature account I just discussed is meant to characterize the paradigm cases of exploratory experiment in the papers by Franklin-Hall, Burian, O'Malley, Elliott, and others, it does a fine job. To the extent that it is meant to capture the entire range of experimental inquiry which is not hypothesis-testing, this account is inadequate. The Lenski experiment (see Chapter 2) is a great example of an occupant of the middle ground: It is not a case of classic hypothesis-testing, but it does not exhibit the paradigm features of exploratory experiment.

At first pass, the point about being theory-informed rather than theory-driven nicely captures what is going on in the Lenski experiment. While the experiment was not motivated by the desire to answer a particular question, it was strongly influenced, informed, and guided by theory. At the outset of the experiment, theoretical background in experimental evolution studies and population genetics motivated the researchers to design this new way to study long-term evolutionary dynamics, and learn from watching them unfold in real time in the laboratory. Over the course of the experiment, more specific theoretical backgrounds from both evolution and ecology have come to bear in learning from the data being generated.

The other two features of paradigm exploratory experiments, lack of solid theoretical background and wide instrumentation, do not characterize the Lenski experiment. Unlike metagenomics, nanotoxicology research, and other cases of exploratory experiment discussed above, the Lenski experiment was not designed to fill an analogous identified wide gap in background theory in the relevant area of inquiry.[16] The experiment operates against the well-established background of evolutionary theory in general, and population genetics in particular. Furthermore, researchers are not gathering data in the massively parallel or high-throughput way characteristic of wide instrumentation.

It is worth noting two points before moving on. First, in saying that the standard account of exploratory experiments does not capture what is going on in the Lenski experiment and thus does not capture what is going on in all non-hypothesis-testing experimental inquiry, I am not arguing that it does a bad job of capturing what it aims to capture. Most of the papers on exploratory experiment focus on characterizing paradigm cases, and their account does a fine job of that. My claim is that, to the extent that the exploratory experiment account in intended to go beyond those cases and capture all non-theory-driven experiments, as some imply (see discussion of Elliott's view below), it is not doing the trick.

The second point to note is that all three of the features I discussed above as applying or failing to apply to the Lenski experiment should be thought of as matters of

[16] Of course, if we had absolutely gap-free background knowledge in any area of science, we would not be worried about gathering more information. It is helpful to distinguish between *quantitative* and *qualitative* gaps in background knowledge. I think it's fair to say that exploratory experiments are concerned with filling qualitative gaps. The difference is between, in the quantitative-gap case, wanting to learn more or get better information about areas where we already know quite a bit (like in the case of better understanding particulars of the dynamics of evolutionary change, informed by evolutionary theory and population genetics), versus, in the qualitative-gap case, wanting to carve out new kinds of sources of theory-generating information in areas where we know very little (like in the case of the toxicology of nanoparticles).

degree, not discrete binary characteristics. An experiment can be theory-driven or theory-informed in a number of different ways and to different degrees, as I discuss in more detail below. Similarly, while there is certainly a qualitative difference between the breadth and throughput of data collection in DNA microarray studies versus in the Lenski experiment, there is no obvious line that allows categorization of every experimental protocol as using only "wide" high-throughput techniques versus traditional low-throughput ones. Finally, while there is a clear difference between the relatively new and patchy theoretical background of nanotoxicology and the established and rich bodies of evolutionary theory and population genetics, there are no obvious criteria for when a theoretical area crosses the threshold from new and underdeveloped to established and well-developed.

I have shown how exploratory experiments are distinguished from classic hypothesis-testing experiments, and how neither picture adequately captures what is going on in the Lenski experiment. There is a conceptual space of experimental inquiry, in which hypothesis-testing and exploration occupy two parts, but not the whole space. The idea that such a space exists is not new. In an early paper on exploratory experiments, Steinle writes:

> The terms ['hypothesis driven' versus 'exploratory' experimentation] do not refer exclusively to specific experimental procedures, rather they indicate a whole range of procedures. In both cases more detailed distinctions can and should be made for further investigation. My claim is not that all experimentation should be subsumed under these two types. There may be experimental procedures of still another character. What I do claim, however, is that my distinction covers some essential aspects of experimentation in scientific research. (1997, p.S69)

Since Steinle's paper the literature on exploratory experiments has focused on fleshing out the picture of exploratory experiments, and little has been said about these "experimental procedures of another character." But this view of a whole range of

procedures has been developed in various ways. The general consensus in the literature on exploratory experiments is that their relationship to hypothesis-testing experiments should be thought of as a continuum, not a dichotomy. While some gestures have been made in the direction of fleshing out what this continuum looks like, there is a lot of work left to do.

Before moving on to propose some further steps in that direction, I should make a point about terminology. As indicated above, there is some confusion in the literature on exploratory experiments as to whether they should be understood as constituting the entire range of non-hypothesis-testing experiments (as Elliott says), or some subset of that range (as Steinle implies). We need a better sense of what non-hypothesis-testing experiments are up to, and the kinds of things that people have talked about as paradigm cases of exploratory experiment do not capture everything important there. I do not want to get hung up on labels; if we end up calling all non-hypothesis-testing experiments exploratory, that is fine, and if we end up calling just a subset of them truly exploratory, that is fine too. My key point is that we need to get clearer about that *whole* space, comprising the paradigm cases and everything in between. The literature on exploratory experiments seems to be gesturing at the whole space with their negative definition, but then talking about only a portion of it with their case-based accounts.

Section 3.3 discusses others' work toward developing the view that there is a range of kinds of experimental inquiry that go beyond the hypothesis-testing paradigm. They have made a good start, but none of their accounts are clear enough about what this space of experimental inquiry actually looks like. This is by no means an easy task, as many different kinds of questions go into thinking about what it is to be an experiment and how experiments fit in to the larger picture of scientific inquiry.

3.3. Spaces of Experimental Inquiry

Even before the literature on exploratory experiments took off, people were starting to push back against the traditional equation of experimentation with hypothesis testing. In his paper "Theory and Experiment in Evolutionary Biology," Brandon (1994) proposed a two-dimensional conceptual "space of experimentality." His overall claim is that, while hypothesis-testing is the paradigm kind of experimentation, scientists can engage in activities that do not exactly fit this paradigm and still be rightly called experimental. His starting point for choosing the dimensions of this space is looking at the relationship between experiment and two scientific activities it is contrasted with: observation and description. From these contrasts, Brandon draws out two methodological questions relevant to determining if a case of inquiry is experimental: First, does it manipulate or not? Second, does it test a hypothesis or measure a parameter?

He argues that if we think of these two dimensions (manipulating versus not and hypothesis-testing versus measuring) as strictly dichotomous, we run into trouble. Manipulative tests of hypotheses are experiments, but there is no obvious reason why manipulative parameter measurement should not count as experimentation. Brandon sets aside this concern by arguing that both dimensions are continuous. The extent to which a researcher manipulates nature (or not) in a given case of inquiry is a matter of degree. Similarly, the same study can be engaged in both hypothesis-testing and measurement. As an example he discusses studies of the strength of selection in natural populations, which often proceed as follows: identify types (like phenotypes or genotypes) in the population, measure some component(s) of their respective fitnesses, and use these data to (1) test the hypothesis that natural selection is acting on the population, and (2) measure the strength

of selection on the different types. Brandon points out that whether this kind of study is thought of as primarily a case of hypothesis testing or parameter measurement depends not on some intrinsic feature of the study, but on how the researcher in question sees it and describes it, or how the relevant scientific community judges it. Furthermore, "[a]fter the fact, one can always recast a parameter measurement as a test of the hypothesis that the parameter takes the value that we have just observed" (Brandon 1994, p. 65). Thus, we cannot always distinguish cases of scientific inquiry into the mutually exclusive categories of hypothesis testing versus description, and calling only the former experiment would seem arbitrary.

Brandon's paper came before the literature discussing exploratory experiment as such, but it is an important early contribution to thinking about experimental inquiry as a space that includes, but is not limited to, hypothesis testing. In the subsequent literature on exploratory experiment, several papers have promoted the view that hypothesis testing versus exploration is not a black-and-white dichotomy. One is by Waters (2007), who briefly but provocatively discusses how the line between hypothesis testing and exploratory experiment is not sharp. He says that thinking about the difference as a simple dichotomy is misleading, and "[t]he fact that theory plays a multiplicity of roles in theory-driven research indicates that the difference between exploratory experimentation and theory-directed experimentation may involve multiple dimensions" (2007, p. 6). Waters does not say what exactly these dimensions might be, but his discussion indicates that they (or at least some of them) should focus on the relationship between theory and experiment.

The most extensive work on fleshing out the space between theory-driven and exploratory experiments is in Elliott's paper "Varieties of Exploratory Experimentation in

Nanotoxicology" (2007). Elliott defines exploratory experimentation negatively as "the full range of experimental activities that do not involve theory testing (and that have therefore been neglected in previous philosophical literature)" (p. 10). He acknowledges that depending on how broadly or narrowly one understands exploratory experiments, one could think of them as occupying this entire range, or a subset of it. Ultimately his concern is with mapping out this space, not with precisely which part of it represents the exploratory part. Thus, his project is a starting point for the same project I am interested in: mapping the conceptual space of experimental inquiry in a way that includes the space occupied by neither classic hypothesis-testing nor the paradigm examples of exploratory experiment discussed in the literature.

Elliott's proposed taxonomy distinguishes kinds of exploratory experimentation along three dimensions:

1. the positive aims of experimental inquiry;
2. the role of theory in experiment; and
3. the methods and strategies used for varying experimental parameters.

He says that we can arrange varieties of exploratory experiment according to their position along each of these dimensions, which are continuous, not discrete. So far, so good; but when it comes to actually describing what the overall space looks like, his account gets confusing. First, he does not say how points along any given dimension are supposed to relate to each other, or what sort of scale is represented on each dimension. Second, he says at the outset that the dimensions are continuous, but the way he describes at least two of them (dimensions (1) and (3)) makes them sound discrete.

Elliott's examples of different values experiments could take on dimension (1) (positive aims of experimental inquiry) include identifying regularities between variables,

characterizing particular entities in time and space, and developing new experimental techniques or instrumentation. This makes this first dimension sound like a qualitative dimension with discrete values. Regarding dimension (2), the role of theory in experiment, he focuses on the point that experiments can be more or less heavily influenced by theory; this sounds more like a quantitative and continuous dimension. But his examples of particular values on this dimension go beyond quantifying how much influence theory has, and sound more like describing qualitatively distinct kinds of roles. His examples of values along this second dimension include drawing on some degree of theoretical background;[17] undertaking experiment explicitly to fill a gap or resolve anomalies in theory by "collecting a wide range of data in hopes of determining how, if at all, a particular theory has gone wrong or how it applies in a somewhat new context" (Elliott 2007, p. 13); and "instructions or strategies for exploration actually play[ing] something like the traditional role of 'theory' in a particular domain" (p. 14), in other words, the role of theory in experiment in this last instance is that of "being constituted by exploratory projects or strategies".[18] The difference between these three examples of "locations" on Elliott's second dimension does not look like one of scale; they

[17] Here Elliott references Franklin-Hall's distinction between theoretical background and local theory, underlining the idea I discussed above that theoretical background, but not local theory, informs exploratory experiment.

[18] It is not very clear what exactly Elliott means by this. One possibility is that this reduces to saying theory plays no role, in which case it could still be a point on this one-dimensional quantitative spectrum (that is, its role is as minimal as possible). But it seems like he is saying something different, like this: Within certain frameworks of scientific inquiry, a role that would normally be filled with theoretical background is filled instead by some commitment to identifying interrelations among results from exploratory experiments. He gives this example, drawing on Waters' (2007) discussion of experimentation in classical genetics:

> … Waters argues that classical genetics from the 1920's to the 1940's included three major cognitive elements: pools of special knowledge, patterns of explanation, and patterns of investigation. The third element, "patterns of investigation," consisted of exploratory strategies that structured research. According to Waters, researchers guided by these strategies designed series of experiments such that explanations of the results, explanations which typically appealed to the transmission theory of inheritance, would reveal information about one or another biological process (and often processes not related to the transmission theory). (Elliott 2007, p. 13)

cannot be thought of as three points in a spectrum from less to more heavily influenced by theory as he indicated in the beginning of his discussion of this dimension. Rather, they look like three different kinds of relationships between theory and experiment. This distinction between qualitative and quantitative aspects of the role of theory in experiment—that is, what kind of role theory plays and how strongly it plays it—is important, but talking about both at once leads to an overall confusing picture of what this single dimension in Elliott's space is supposed to look like.

Elliott's third dimension, methods and strategies used for varying experimental parameters, looks like the first in that his description of it implies a qualitative dimension with discrete values. Candidate values include "systematically altering the features of an experiment in order to uncover regularities," "study[ing] a phenomenon using as many tools and techniques as possible so as to understand it more fully and to gain more solid epistemic access to it," and "working as a community to collect experimental results under a wide variety of conditions" (pp. 14–15). These do not sound mutually exclusive. Furthermore, they are put vaguely enough that it is unclear why we should think of them as features particular to exploratory experiment, rather than potential features of any experiment. More concrete values on this third dimension include using high-throughput (or "wide") instrumentation as opposed to traditional instrumentation. Again, it is not clear how this sits with respect to the other three options. High-throughput instrumentation is a tool and technique which can be used to study a phenomenon to understand it more fully (the second cited example of a value on this dimension, above). So why should we think of these as two separate points on a dimension of methods for varying experimental parameters?

Missing from Elliott's account is a sense of what the dimensions, and thus the space as a whole, actually look like. For instance, what is the scale that makes something fall at one end versus the other of the methods and strategies dimension? He intends this three-dimensional space to be a functional classification scheme: "To place a particular sort of exploratory experimentation within the taxonomy, one can identify where it fits along each of the three dimensions" (2007, p. 15). It is not clear how the example characteristics mentioned above (altering experimental features, using many tools and techniques, and working as a community) constitute a linear pattern that we could think of as a one-dimensional continuum.

To be fair, Elliott is quick to point out that there is more work to be done to improve his taxonomy of experiments (2007, p. 18), and furthermore, it is an open empirical question whether or not this space can actually be mapped. By defining this space at all, he has achieved his goal of providing a "more complete and systematic account [than has been previously given] of the ways in which different forms of [exploratory experiment] compare with each other" (p. 4). Two good starting points for a better account of this overall space would be (1) being clearer about which features of experiments can be talked about qualitatively versus quantitatively, that is, distinguishing "how-much" questions from "what-kind" questions, and (2) distinguishing the dimensions in ways that would allow natural ranking along continua like less-to-more, weaker-to-stronger, etc.

One question that immediately comes to mind in thinking about how to give an improved account of the whole territory of experimental inquiry—hypothesis-testing, paradigm cases of exploration, and everything in between—is: What is the point? The list of things we could say to describe an experiment seems endless: in what manner and how

strongly it relates to theory, which theory or theories it relates to, what background information it draws on beyond theoretical content, what information it generates or purports to generate, how the experimental system relates to the researcher's target(s) in the natural world, the nature of those target(s), how many parameters are varied in the experiment, how they are varied, details of the protocol… this list could go on and on. A crucial question in accounting for the "space of experimentality" is to decide, and justify, which features of experimental inquiry matter to defining this space.

In Brandon's and Elliott's efforts to define such a space, we saw two different ways to think about this. Brandon started with trying to distinguish experiment from two of its contrast classes, observation and description; Elliott started with trying to capture what distinguishes non-hypothesis-testing experiments from those that test hypotheses. Here is a summary of their portrayals of the space:

Brandon's two dimensions:

(a) hypothesis-testing/not

(b) manipulating/not

Elliott's three dimensions:

(c) positive aim

(d) role of theory

(e) methods for varying parameters

Collectively, these dimensions address three different aspects of experimental inquiry: the experiment-theory relationship (a, c and d), the experimental system and protocol (b and e), and the experimenter's intentions and motivations (c).

I take (c) to involve two of these aspects because some of the examples Elliott gives for this dimension seem to explicitly involve theory (such as "identifying major

regularities between variables and developing a corresponding conceptual scheme"), and others do not (such as "varying multiple design elements of an instrument to determine which design will be most effective") (Elliott 2007, p. 11). Perhaps everything Elliott talks about as part of this first dimension of his space could be folded into the other two. Saying that the positive aim of a given experiment is to identify regularities and develop a conceptual scheme implies that theory plays a particular role in the experiment, namely, as its aim: The point of this experiment is not to test theory, but to generate it. Likewise, saying that the positive aim of a given experiment is to vary the instrumentation to figure out how to use it most effectively implies the use of one method for varying parameters (or, perhaps setting the stage for effective parameter-varying).

In the rest of this chapter, I build on both Brandon's and Elliott's accounts as starting points for further clarifying a key piece of the conceptual space of experimental inquiry: the relationship between experiment and theory. This tracks both Brandon's dimension about hypothesis-testing and what I think Elliott had in mind with his dimension about "positive aims." It seems right to say that anything we could say about the positive aims of an experiment could be talked about in terms of the experiment-theory relationship, features of the experimental system and protocol itself, or the experiment-world relationship (that is, what the experimental object is meant to tell us about the target).

I ultimately think that a complete account of the conceptual space of experimental inquiry should address all three of these latter aspects. In the remainder of this chapter I begin to address the first, the experiment-theory relationship. I give a preliminary account of a more nuanced and realistic way to approach thinking about experiment-theory relationships.

3.4. Theories Can Play Many Roles in Experiments

The previous accounts discussed above tend to talk as if a single body of theoretical background plays one role at the outset of each experiment (either as driver or informer). This is misleading. On the hypothesis-testing view, that picture looks something like this:

THEORY ⟶ **HYPOTHESIS** ⟶ **EXPERIMENT**

As Hempel, Hansen, and others describe it, researchers operate within the context of particular theoretical backgrounds, and specific hypotheses are generated from those backgrounds; these motivate the researchers to design and undertake experiments to test the hypotheses. The exploratory experiment picture is different. There, experiments are informed but not driven by theoretical background, and theory (or the generation of new information to eventually fortify theory) follows from it as an aim. As Elliott (2007) describes it, theory might play little role in the design or guidance of exploratory experiments, but it plays a role in their motivation, namely, they aim at developing new theoretical ideas and concepts. So, the picture there looks more like this:[19]

THEORY ⇢ **EXPERIMENT** ⟶ **(NEW INFORMATION, NEW THEORY...)**

In both accounts of experimentation, the picture focuses on one body of background theory playing one role in an experiment, and that role being played at the experiment's temporal outset: either as provider of a specific hypothesis for the experiment to test, or as provider of noteworthy gaps to be filled.

[19] The grey/dotted line is meant to indicate that that theory at the outset has obvious holes which need to be filled in.

This is misleading because multiple, diverse theoretical backgrounds commonly come to bear in individual experiments. Theories can also come to bear differently at different points in an experiment. They can direct, guide, influence, inform, and result from experiments all at once; new theories can come into play over the course of the experiment which researchers were not drawing on at its outset; they can also fail to play much of a role at all. I think a good starting point for a better understanding of the conceptual space of experimental inquiry is getting rid of the notion that "the role of theory in experiment" could be a single dimension in that space. Different questions need to be asked here, and they should be asked about the roles of theories in experiments. These include: Which theoretical backgrounds are coming to bear in the experiment? What roles are they playing, and how strongly are they playing them (for instance, does addressing a particular aspect of theoretical background generate the entire motivation for doing the experiment, or is it one influence among many)? When do they come to bear over the timeline of the experiment, from its original motivation to its design to its implementation to its end? Is a hypothesis being tested at the outset? Are hypotheses being generated along the way or retroactively?

The following list is a starting point for separating out these sorts of questions in the context of a given case of experimental inquiry. I do not think that this list is exhaustive, but it is a start toward being clearer about the different roles that exist. Here are three different questions we could ask about the roles of theories in experiment:

1. **Origin:** What role does theory play in an experiment's *origin* or *motivation*? Was the decision to do the experiment motivated by desire to test a particular hypothesis generated by some particular theory? Was it motivated by identifying a particular area of theory that lacks empirical

validation, or a problematic conflict between the predictions of rival theories that needs to be settled?

2. **Direction:** What role does theory play in the experiment's *direction* or *deciding what to do next*? Once the decision to do the experiment has been made, what theoretical background, or specific theoretical guidance, goes into figuring out what to do next if things are not going as planned, or modifying the original course of action if unanticipated results or surprising dynamics emerge?
3. **End:** What are the experimenter's ultimate *aims* with respect to theory? This can take a number of forms, including answering specific questions, resolving anomalies, adjudicating among rival theories, generating information to fill gaps in theory, or laying groundwork for new theoretical backgrounds.

When I say that previous attempts to map the space of experimental inquiry have focused only on theory's role at the outset of an experiment, I do not mean that they have focused on only the first kind of role I distinguish above, that of **origin** or motivation. I mean that they have made it sound like experimenters establish a particular goal with respect to theory at the temporal outset of an experiment—to test a hypothesis in the classic case, or to generate more or better theory in the exploratory case—and that this goal determines and constitutes *the* role of theory in that experiment. My three-part distinction is intended to make explicit that this is often not how it works. Others have pointed to the distinction between what I am calling theory's roles as **origin**, **direction** and **end**, but they tend to talk as if one theory plays one of these roles. Elliott, for example, in his discussion of his second dimension (the role of theory in experiment), says that "[i]t is important to

remember… that although Steinle's preferred form of EE [exploratory experiment] includes very little role for theory in the design or guidance of EE, the goal of the activity is to develop new theoretical ideas and concepts" (2007, p. 12). However, he thinks of the role of theory in experiment as classifiable along a single dimension, which implies that an experiment can have a value on this dimension representing *the* role theory plays in it. The point of my account is to underline that often it does not make sense to talk of *the role* that *theory* plays in an experiment, but rather the *roles* that *theories* play in an experiment.

On the classic view of experiments as hypothesis testing, these three roles are all filled by the same body of theory. Consider again the antimutator example discussed above. In this case, the relevant theoretical background concerns genetic mutations: general theory about their evolution and mechanisms combined with Drake's particular arguments against the probability of general antimutators. This background, prompted by details about the previous results being scrutinized in the experiment, led to the question: Is *dnaE911* a general antimutator in strains of *E. coli* other than strain *K12*? This question motivated the experiment in the first place (**origin**), drove its design (**direction**), and its answer constituted the aim and final upshot of the experiment (**end**).

The Lenski experiment, in contrast, is a case where it is wrong to think about the theory-experiment relationship in terms of one theoretical background playing one role. There, evolutionary theory and population genetics were in the background from the outset; these played an **origin** role which is best classified as informing, but not driving. The experimenters set out to learn about the dynamics of evolutionary change in a novel and relatively open-ended way; their aim was not to confirm or reject a specific hypothesis about those dynamics. Over the course of the experiment, elements of theoretical background came to bear which were different from those that served as the experiment's

motivation, either because they were much more specific (as in the case of developing theory on the evolution of mutation rates (Sniegowsk, Gerrish, & Lenski 1997)), or because they branched into areas outside of evolutionary theory (as in the case of ecological theory about multiple populations' stable coexistence (Turner, Souza, & Lenski 1996)). These theoretical areas came into play as both **direction** (for example, by motivating measurement of new parameters and other additions to the protocol) and **ends**. Thus, with the three-part distinction above, we can say something different about each role of theory, and about distinct theoretical backgrounds playing each kind of role within the context of a single case of experimental inquiry.

For the paradigm cases of exploratory experiment, finally, we might have relatively little to say about the first or second roles theories can play, **origin** and **direction**. In some cases there might be no particular background theory playing a notable role in the experiment's origin or direction; the experiment might be undertaken wholly or primarily to produce new information. In this case, we could focus on the role of theory primarily in the third sense, the **end**.

I criticized Elliott above for implying that cases of experiments could be situated at particular locations along his dimensions, without being clear about what those dimensions look like exactly. So far, my proposal doesn't seem to fare much better on this count. But my proposal is not that **origin**, **direction**, and **end** are three linear dimensions along which cases of experiment can be neatly situated. Rather, it is that to understand the conceptual space of experimental inquiry we need to understand these as three different roles theories can play in experiments. If we like we can think of these as three "dimensions" in a space, but visualizing how experiments fit in this space is going to get complicated. It is more than just a matter of situating cases in a three-dimensional space.

This is because regarding each of the three roles we should ask some further questions about the bodies of theory in question. In particular (and this is not necessarily exhaustive):

1. A general question: Are any two, or all three, of the roles shared by one body of theory, or are different bodies of theory coming into play for each role? (This is a key question which will set most cases of classic strict hypothesis-testing apart from the rest).
2. Two more specific questions: Are these bodies of theory theoretical background or local theory, in Franklin-Hall's sense? And, even more specifically, what are they?

Even if we ignore the specific questions, it will get complicated to visualize how particular cases could fit in this space.[20] But focusing just on the three dimensions and the general question, we can see that classic hypothesis-testing experiments and the paradigm cases of exploratory experiments, respectively, will occupy particular regions in the space. And that leaves plenty of unoccupied territory as well, what I've been calling the space in between. That is where cases like the Lenski experiment, and many other examples (not just from experimental evolution but across science) will fall.

Classic hypothesis-testing cases will involve a single body of theory playing specific roles in **origin** and **end**, namely, generating a specific hypothesis to test and either rejecting or failing to reject that hypothesis, respectively. That same body of theory will play the **direction** role; this can be done in a number of ways, but the key distinguishing feature of classic hypothesis-testing in this space is that a single theory occupies all three

[20] And remember, this is not meant to be a replacement for the entire conceptual space of experimental inquiry—here I am focusing just on the aspects of the space which regard the experiment-theory relationship. Adding in further important aspects, like salient features of the experimental system and protocol and experiment-world relationship, would make things more complicated.

roles, and in a very particular sense with regard to **origin** and **end.** For example, in Section 1.1 I discussed the Meselsohn-Stahl experiment, where the body of theory in question regarded the mechanism of DNA replication. A specific hypothesis was on the line, the hypothesis that DNA replication is semiconservative, and the experiment was designed with the aim of achieving an intervention which would test that hypothesis and rule out alternative possible mechanisms of DNA replication.

The paradigm cases of exploratory experiments will look quite different in this space. In these cases at least one body of theory will play a key role in **end**: The experimenter's aim is to fill significant gaps in that theory, augment it, or perhaps develop it from scratch. Theory plays a weaker role in **origin** and **direction**, and the role it does play there is typically occupied by a different body of theory than that which is the experimenter's end. For example, in the case of fMRI studies: We know little about how particular patterns of brain activity correlate with particular cognitive states, and the goal of this research is to collect huge amounts of data, generate new hypotheses to test, and eventually develop a robust body of theory about these connections. A different set of bodies of theory than the latter—including theory about how the brain works in general, and how to build the appropriate scanners—informs fMRI experiments in both the senses I'm calling **origin** and **direction**.

So classic hypothesis-testing and exploratory experiments occupy two different parts of this conceptual space, and together they do not occupy the whole space. How would we situate the Lenski experiment in this space? There are a number of ways to think about this; doing so brings us back to the question I raised and set aside in Section 2.1.5, namely, what is the experiment? One way to think about it would be the entire 27+ years of propagation of the *E. coli* populations; another way would be to treat each individual

paper that has been published as a self-contained experiment to fit in the conceptual space. There is not right answer here. The important point is that either way of thinking about it will result in something which occupies neither the classic hypothesis-testing or paradigm exploratory experiment portions of the space.

Much more work is needed to develop an account of the conceptual space of experimental inquiry which fully characterizes the space in between classic hypothesis-testing and paradigm cases of exploration. In future work I will develop further an account of what this space looks like and how to situate particular cases in it. My aim in this chapter was just to first show that previous accounts did not do the whole job, and second, lay the groundwork for a key piece of doing the job better: a more realistic picture of the complex relationship between experiments and theories.

4. Experiments and Simulations[21]

In this chapter I respond to a debate about experiments versus simulations which involves arguing that experiments have epistemic privilege over simulations. Experiments are thought to have two particular virtues which give them this privileged status. First, they generate greater inferential power, or external validity: Experimenters are in a better position to make valid claims about their targets of inquiry in the natural world. Second, experiments are a superior (or, the only) source of surprises or genuinely novel insights. I will argue that both of these claims are false as generalizations across science. All of scientific inquiry involves engaging with some object of study—a model, a physical system in the laboratory or field, or a combination of these—to learn about some target of inquiry. The methodological distinction between experiment and simulation is certainly important for making judgments about epistemic value. But I will argue that this is the case only in a context-sensitive way, not as a generalization across science. The experiment/simulation distinction should not be used as a basis for in-principle judgments about epistemic value. Whether we are better off studying a phenomenon in the world by interacting with experimental systems or computer simulations depends on a complex of factors—including the kind of question we are asking and what we are asking it about—and not on some absolute assessment of the primacy of one kind of scientific inquiry over another.

[21] Parts of this chapter are published in Parke (2014b).

4.1. Some Preliminary Points About Experiments and Simulations

Scientific practice in the twenty-first century is increasingly blurring the lines between experiment and simulation. While it was once common for individual scientists, laboratories, or even entire subfields to focus on only one of these methodologies, experimental and computational methods are now increasingly combined. This has led to new ways to do science, as well as opportunities to reexamine the roles that experiment and simulation play in scientific inquiry, and their changing natures in practice. These trends are reflected in increased attention from philosophers of science to experiment (see discussion in Chapters 1 and 3), simulation (e.g., Humphreys 2004; Weisberg 2013; Winsberg 2010), and their methodological and epistemic points of convergence and contrast (e.g., Barberousse, Franceschelli, & Imbert 2008; Guala 2002; Morgan 2005; Morrison 2009; Parke 2014b; Parker 2009; Peck 2004; Peschard 2012; Winsberg 2009).

There is a pervasive view among philosophers and historians of science, and scientists themselves, that experiments have epistemic privilege over simulations. That is, they allow us to make better inferences about the natural world, or generate more reliable and trustworthy scientific knowledge. Simulations are often talked about as a fallback, something a scientist should do only when an experiment would be too cost-prohibitive or otherwise impractical. A number of people have recently put this kind of idea in writing. For example: "[Simulation's] utility is debated and some ecologists and evolutionary biologists view it with suspicion and even contempt" (Peck 2004, p. 530); "simulations are supposed to be somehow less fertile than experiments for the production of scientific knowledge" (Guala 2002, p. 4); and "the intuition of [non-economic] sciences

and philosophy is that experiment is a more reliable guide to scientific knowledge" (Morgan 2005, p. 324).

Experiments are thought to have two particular virtues which give them a privileged status over simulations. The first is that they generate greater inferential power, or external validity: Experimenters are in a better position to make valid inferences from their objects of study to their targets of inquiry in the natural world. The second is that experiments are a superior (or the only) source of surprises or novel insights. I will argue that both of these claims are mistaken as generalizations. To the extent that the difference between experiment and simulation carries epistemic weight, this weight is context-sensitive; the mere fact that a case of scientific inquiry counts as an experiment or a simulation is no indication of its epistemic value. There is a lot of important work to be done in understanding what grounds and validates inferences from objects of study to targets of inquiry in the natural world (I discuss this further in Chapter 5). Focusing on whether to classify cases as simulations or experiments, per se, muddies the waters of that task.

In arguing that there is no in-principle difference in epistemic value between experiments and simulations across science, I am not endorsing the positive claim that experiments and simulations are epistemically on a par across science. I am not challenging the idea that, within the context of a particular research area, there might be good reasons to think that experiments have epistemic privilege over simulations—or vice versa. These might be historical reasons, like experiments having a particularly good track record compared to simulations in that research area, or more principled reasons, like lack of enough data to build computer models which we could trust anywhere near as much as we trust experiments (this is the case, for example, in clinical trials of new medications).

There might also be good reasons in particular research areas to think that simulations have epistemic privilege over experiments, for example, areas where experiments are impossible like studies of long-term climate change. But these are both context-sensitive points about particular contexts of inquiry, not about experimental versus simulation methodology across science.

Furthermore, I am not challenging the idea that experiments have priority over computer simulations in the greater picture of what fundamentally grounds scientific knowledge. We know how to do good computer simulations precisely because we have gained knowledge about the world through observation and experiment. In the grand scheme of things, empirical data is fundamental for answering scientific questions about the natural world. Theory plays an important role in simulations designed to teach us about particular targets in the natural world; as Winsberg nicely puts it, "In such contexts [i.e., most simulations though especially in the physical sciences], simulation is only possible precisely because good theories, well confirmed by a history of experiments, exist to underwrite them" (2010, p. 29). However, people often seem to blur the lines between this general empiricist claim and a separate issue, namely, which methodology, now, will generate better scientific knowledge. This is the target of my objection: The kind of thinking that goes into claims about experiments' superior position in a hierarchy of scientific methodologies, like those cited above from Guala, Peck, and Morgan, or claims that a case of research is less epistemically valuable because "it's just a simulation."

Much of the literature on experiment versus simulation focuses on computer simulations, studies of computational models with some dynamic temporal element. But there is another, broader understanding of 'simulation' where the object of study in question could be any kind of model: mathematical, computational, or concrete

(physical). Simulations in this broader sense are taken to include studies of computer models, model organisms in laboratories, and model airplanes in wind tunnels. Discussions of the relationship between experiment and simulation in the literature have focused sometimes on computer simulation (Humphreys 2004; E. F. Keller 2003; Morrison 2009; Parker 2009; Peck 2004; Winsberg 2003) and sometimes on simulation in the broader sense (Guala 2002; Morgan 2005, p. 320; Winsberg 2009, 2010). I highlight this contrast upfront because being clear about which sense of 'simulation' is at stake matters. In particular, I do not believe that there are interesting or important methodological or epistemic distinctions to be drawn between experiments and broad-sense simulations. I will say more about why in the following sections.

4.1.1. Simulations or Models?

Before moving on to discuss the arguments for the epistemic superiority of experiments, I should address a question that might come up: Why am I talking about simulation rather than modeling? The short answer is that I am effectively talking about both, because I take simulation to be the practice of studying models. But this sort of question might arise because people have talked about the relationship between simulation and modeling in different ways. The most common way to think of this relationship, and the way I think about it, is that a simulation is a study of a model, generally involving some dynamic temporal element (studying what happens to the model over time). Most people talking about simulations understand them in roughly this way. Peschard, for example, exemplifies this view when she defines a simulation as "the manipulation of a putative model of the target system" (2012, p. 12). Weisberg (2013) says that simulations are ways to compute the behavior of a model using a particular set of initial conditions, contrasting

simulation with mathematical analysis as methods for investigating models. Winsberg defines simulation as the "comprehensive process of building, running, and inferring from computational models" (2003, p. 107).

There can be confusion here, though, because people understand 'simulation' in different ways, sometimes contrasting it with modeling. For example, a recent paper on simulation talks about it as follows: "Recent years have seen a developing discussion on the role and epistemology of simulation in modern scientific practice, as a subject worthy of its own attention, distinct from experimentation and modeling" (MacLeod & Nersessian 2013, p. 533). Others talk about simulation as a particular *kind of* model, rather than an activity which one *does with* (or to) a model. One way this gets fleshed out is by saying that simulations are a particularly realistic kind of model, which involves an explicit dynamic element of modeling a system's states over time. Peck and Lenhard, for example, both seem to endorse this kind of view when they talk about "simulation models." Peck (2004) contrasts these with "simple models," by which he means non-computational mathematical models, such as the basic models of population genetics. He separates simulations from other kinds of models, and says that simulations should actually be thought of as experiments, while (simple) models are a contrast class to experiments:

> The kinds of experiment done with the simulation model give insight into future data-gathering efforts, test hypotheses that would be impossible to test otherwise and inform researchers about the implications of theoretical insights contained in the causal story that the model represents. Simulation is another experimental system with which to explore theories about how the real world works, using an artificial world that researchers can control. (Peck 2004, p. 533)

In a related vein, Lenhard (2007) talks about simulations as a special kind of model, rather than something one does with or to models, though unlike Peck he thinks that simulations are in a different category from experiments. He says that philosophers of

science have viewed the models that do the "real work" in science as the continuous mathematical functions characteristic of theoretical models, and simulations as just discretized versions of those functions. Simulations, he says, are formulated in a different way from other models, so deserve their own classification; it is incorrect to assume that there is a prior model and the simulation is just a "realization" of it. He argues that "simulation models," in fact, should be thought of as a separate category of model:

> Simulation models do not just apply the brute force of the computer in order to squeeze results out of [theoretical models consisting of continuous mathematical functions]. Instead, they require their own new kind of modeling… This type of discrete model defines the spectrum of potential models anew, motivated by the specific requirements of the computer. The reason for this is that the generative mechanism selected to imitate a certain dynamic has to "run" on the computer; that is, it must not require excessive computing capacity and must, above all, not become unstable, because, for example, discretization or truncation errors will build up. (Lenhard 2007, p. 187)

It is worth noting these different views on the relationship between simulation and modeling, and how these relate in turn to experiments. In any case, I do not have a particular stake in how this definitional issue is ultimately settled, and will set it aside. I am interested here in the methodological and epistemic contrasts people have made between studying experimental systems and studying models, and I am focusing on simulation here, understood as the activity of studying models, because that is the main activity people have contrasted with experimentation. Perhaps it makes sense that this is a cleaner grounds for comparison than experiments versus models, because simulations (understood thus) are about studying models, and experiments are about studying physical systems in a laboratory, in the field, etc., both with the goal of learning about some target system. Construed thus, both methodologies are about manipulating an object to learn about a target.

4.2. The Argument about Inferential Power

Belief in the epistemic privilege of experiments over simulations is often grounded in ideas about their relative inferential power. In particular, the idea is that experiments lead to better inferences about natural systems or phenomena than simulations do (this is sometimes referred to as the issue of external validity). This difference has to do with the relationship between their respective objects of study and targets of inquiry.

Judgments about experiments' privileged status are often driven by the intuition that experimental objects of study have a privileged link to targets in the natural world in virtue of their shared materiality. Winsberg nicely describes this as "the suspicion (or conviction) [that] the experimenter simply has more direct epistemic access to her target than the simulationist does" (2010, p. 55). Morgan (2005) and Guala (2002) have argued that material object–target correspondence is a defining feature of experiments, and that this correspondence is responsible for experiments' advantage over simulations in terms of inferential power. I will call this shared view of theirs the *materiality thesis*.[22] Guala puts it in terms of experiments' objects having "deep, material" correspondence to their targets, while simulations' objects have only "abstract, formal" correspondence to their targets. He says that in virtue of their very design, experimental systems are in a better position to produce external validity given their material correspondence to the outside world:

> The trick is to make sure that the target and the experimental system are similar in most relevant respects, so as to be able to generalise the observed results from the laboratory to the outside world. Experimenters make sure that this is the case by using materials that resemble as closely as possible those of which the parts of the target system are made. (Guala 2002, p. 12)

[22] Harré (2003) also argues for a version of this thesis, though in the context of discussing inferences from experiments, not comparing experiments to simulations.

Morgan uses a different set of concepts and terms to argue for essentially the same view; she puts it in terms of experiments' objects "replicating" or "reproducing" parts of the world, while simulations' objects only "represent" parts of the world. On her view, what experimental systems versus models are made of plays a crucial role in the epistemic consequences of experimenting versus simulating: Experiments have greater potential to generate valid inferences about the world, precisely because experimental objects are "made of the same stuff as the real world" (Morgan 2005, p. 322). In other words, material correspondence implies greater inferential power.

These authors cite examples of experiments in economics and psychology to support their points. There, experimental objects of study (humans) are taken to be the same kind of thing in the laboratory and outside the laboratory. Human subjects are examined in laboratory scenarios which aim to be realistic, in order to elucidate market dynamics of interest. Their point about material correspondence is supposed to be that the experimental object and target are literally made of the same stuff: real humans and their real participation in market dynamics. A model or simulation could represent these dynamics and the various causal relationships involved, but it would be lacking the material correspondence to the target system present in the experimental setup. As Morgan puts it, "[T]he fact that the same materials are in the experiment and the world makes inferences to the world possible if not easy… the shared ontology has epistemological implications. We are more justified in claiming to learn something about the world from the experiment because the world and experiment share the same stuff" (2005, p. 323). Thus her endorsement of the materiality thesis: Experimental inquiry gets us closer to the natural world because experimental objects are samples, instantiations, or reproductions of their targets; simulations, in contrast, have as their

objects (only) models of their targets. She concludes that "on grounds of inference, experiment remains the preferable mode of enquiry because ontological equivalence provides epistemological power" (Morgan 2005, p. 326).

There is certainly some truth in the intuition behind the materiality thesis. Experimenting on an actual sample or physical approximation of a target in the natural world is often the best way to get traction on understanding it when we know very little about the target in question, or when the relevant theoretical background is minimal. For example, we have relatively low confidence (today) in our understanding of how new drug cocktails will work in the human body. It makes sense for us to place more confidence (now) in tests on physical proxies, like mice or, even better, human clinical trial volunteers, than in a computer simulation of the human body. But often is not the same as always. Material object–target correspondence is not, and should not be thought of as, necessarily the best route to valid scientific inferences.

4.3. Responses to the Materiality Thesis

In this section I raise two problems for the materiality thesis, in increasing order of severity. The first is a general point about how experiments are designed. The second returns to some examples of laboratory natural selection experiments from Chapter 2 to show why the notion of material correspondence itself, and the ensuing distinction between experiments and broad-sense simulations, is not as straightforward as Guala and Morgan take it to be. This leads to several points: First, we should reject the idea that material object–target correspondence is characteristic of experiments and confers greater

inferential power on them; second, we should stop looking to the experiment/simulation distinction altogether as a basis for making wholesale judgments about inferential power.

I agree with concerns that both Parker (2009) and Winsberg (2009; 2010) have raised about the materiality thesis; in particular, their points that material correspondence does not always entail greater inferential power, and that it is difficult to even make sense of the distinction between material and formal object–target correspondence. I am objecting to the materiality thesis in a different context than theirs, namely, that of arguing that we should do away altogether with relying on the experiment/simulation distinction to tell us anything in principle about epistemic value. Winsberg (2010) seems sympathetic to my conclusion, but I am putting the point more strongly: It is not just that there are exceptions to the generalization that experiments have epistemic privilege, but rather, thinking in terms of such generalizations is the wrong way to approach judging the epistemic value of cases of scientific inquiry.

4.3.1. Experiments Are Not Necessarily Designed To Be Like Particular Targets

To the extent that Guala's and Morgan's attribution of superior inferential power to experiments rests on their being designed to correspond materially to predetermined targets in the world, this account has a problem. Experiments are not always designed this way. Experimenters do not always go in to their experiments with a clear idea of what their targets in the world are "made of," for at least two kinds of reasons: (1) they do not

have a particular target in mind at all,[23] because they focus initially only on designing a particular sort of experimental system and have not yet decided or figured out what kinds of inferences to the outside world they will draw from it, or (2) they have a particular target in mind but have minimal knowledge of its properties. I will discuss both of these reasons in turn.

One kind of case where experimenters do not know what their target is made of is when they do not have a concrete target outside of the experimental system in mind when they design the experiment. Many, perhaps even most, experiments are designed to answer a particular question about the world, and thus designed with a particular target in mind to tell us something about. This is certainly true of the economics examples Morgan discusses. But this need not be the case; the paradigm examples of exploratory experiments discussed in Chapter 3 are cases in point. Sometimes an experimenter's goal is to design an experimental system, study it, and see what happens, or to collect as many data as possible and then figure out how to interpret them and what we can learn from them. The Lenski experiment, while not a paradigm case of exploratory experiment, is another example of an experiment designed without a specific target in mind. The experiment was designed to create an object of study which would offer new kinds of insight on the dynamics of long-term evolution. Its design was driven by evolutionary theory, but it was not designed to test a single hypothesis about the world that came from that body of theory, and it was not designed with a particular target in mind that it was

[23] I should underline that I am using 'target' here, as I do throughout the dissertation, to mean the system of study in the natural world about which it is a researcher's ultimate goal to learn or infer something. When I say they do not have a particular target in mind I mean just that they do not have a particular system fitting this description in mind—not that they have no target in mind in the colloquial sense of the term, that is, that they have no aim.

supposed to tell us something about.[24] Intended routes of inference about the natural world were not built into the experimental setup from the getgo, as Guala says (and Morgan sometimes implies) is a universal feature of how experiments are designed. These came after the system was designed and its dynamics observed over time.

A second kind of case where experimenters do not know in advance what the target is "made of" is when they have no empirical access to it in principle. This is usually not the case in biology, where the targets of interest are natural populations of organisms and their evolutionary trajectories, which we can observe to varying degrees. But in certain sorts of cases experiments aimed at filling in our understanding of long-ago evolutionary history have this feature. An example of this is research on the origin of life. One key aim of experimental work on the origin of life is to demonstrate how certain processes, like spontaneous formation of lipid bilayer cell walls or the encapsulation of information-carrying molecules therein, can take place and therefore might have taken place approximately four billion years ago in the prebiotic soup. The target itself (the set of events, entities, and processes involved in the actual origin of life on Earth) is empirically inaccessible to us for obvious reasons. So the aim of creating trustworthy experimental systems here cannot be to make them as much as possible like the target, in the same way we might think of making a scenario in an economics lab as much as possible like real market dynamics in the world. Rather, the aim is to set up an experimental system, using

---

[24] What exactly is the target, then? It seems too broad to say that the target of the Lenski experiment is the set of all past, present, and future entities subject to evolutionary processes. Likewise, it seems too narrow to say it is the set of all *E. coli* in the world (just because *E. coli* happen to be the organisms in the experimental system) or the set of all bacteria, asexual organisms, etc. The experiment was designed to inform us about evolutionary processes, and the laboratory organisms were chosen for their virtues as experimental subjects, as discussed in Chapter 2, not because they have some special correspondence to a given set of organisms in the world designated as future targets of inference. It makes sense to think about the targets of the Lenski experiment only in the context of particular claims which are made about the world, for example, macroevolutionary trends in the fossil record or pathogenic *E. coli* populations outside of the lab, in the punctuated evolution and high mutation rates discussed in Chapter 2, respectively.

the best theoretical knowledge available, which we have reason to think might teach us something about how the target *might have been.* This same sort of point is true of the historical sciences in general.

I raise these considerations just to point out that judgments about inferential power should not be grounded in experiments being explicitly designed to be "made of the same stuff" as their targets, or to capture aspects of a predetermined and well-characterized portion of the outside world in their material constitution, as Guala and Morgan imply. Inferences about targets in the natural world can be thought of after the experiment has been designed, even after it has been carried out. This is an objection to the way Guala and Morgan talk about the process of experimental inquiry, but it is not really a severe point against the materiality thesis. It is more a point about timing than about material correspondence or lack thereof, per se. My response in the following subsection presents a more serious challenge to the view that material object–target correspondence confers superior epistemic power on experiments.

4.3.2. Material Correspondence: Hard to Evaluate, Not Always the Goal

Recall two cases from the Lenski experiment which I described in Sections 2.1.1 and 2.1.2, high mutation rates and punctuated evolution. In the high mutation rates case, an inference was made from the increased mutation rates observed in three of the Lenski populations in the first 10,000 generations, to a proposed similar mechanism for increased mutation rates observed in populations of *E. coli* and other pathogenic asexual microbes in nature. In the punctuated evolution case, an inference was made from the punctuated morphological evolution observed in the Lenski populations in the first 3,000 generations,

to a proposed mechanism for punctuated equilibrium in macroevolutionary trends in the fossil record, over a hugely greater evolutionary timescale.

An important question is: Are these two different inferences plausible and warranted? In the high mutation rates case, the extrapolation claim seems quite plausible. The researchers are tentatively saying that they saw an increased genomic mutation rate in some of the asexual populations in the lab, and that the phenomenon of mutator hitchhiking could explain increased genomic mutation rates both in those experimental populations and in relevantly similar asexual populations outside of the laboratory. In the punctuated evolution case, on the other hand, the extrapolation claim looks more dubious, because punctuated equilibrium is a hypothesis about extremely long-term morphological evolution spanning speciation events; this is quite different from the increased cell size observed in the experimental populations over 3,000 generations. I will return to this question in Chapter 5. For now, I am going to step back and focus on a more fundamental question: What about this experimental system, and its relationship with the dynamics and populations in the world outside of the laboratory in these two different cases, *would* make these valid extrapolation claims?

The materiality thesis implies that inferential power is proportional to the degree of material correspondence between object and target. To think through what this means exactly, we first need to clarify what this correspondence is supposed to consist in. Morgan and Guala refer to it in various ways: "material correspondence," "material analogy," "ontological equivalence," or being "made of the same stuff" (Guala 2002; Morgan 2005). They must of course mean "made of the same kind of stuff," in the sense that object and target are both instances of the same material type. But there are various ways to interpret what this means. One (uncharitable) way would be "made of the same

material stuff at the most fundamental level." On this interpretation, almost all biology experiments could be said to achieve strict material correspondence, because their objects and targets are made of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur.

Here is a more reasonable interpretation of 'material correspondence': The highest degree of material object-target correspondence would be an identity relation. Certain field experiments might achieve this, for example, when they involve studying every individual in a small, clearly delimited population in order to make inferences about particular features of exactly that population. But this sort of object-target identity is far from the norm in science.

The next level of material correspondence would be an object which is a token of the same relevant ontological category as the target, at a sufficiently fine-grained level for the purpose at hand. When a chemist studies samples of uranium to learn about the properties of uranium in general, they are achieving material object-target correspondence in exactly this sense. The object is a token of the target's type at the finest level of grain with which scientists classify the relevant kinds, chemical elements. Further from this extreme would be studying mice to learn about humans. Both are living organisms, but they are not of the same type at even close to the finest level of grain with which we classify organisms, phylogenetic classes. They are both mammals, but do not belong to the same species or even genus. Even further from this extreme would be studying plastic models of mice to learn about mice.

I take this to be the most plausible way to make sense of the idea of degrees of material correspondence, and will from here on use the term in this sense: in terms of variations in grain of correspondence at the relevant "material" (that is, chemical, physical, biological…) level of categorization. (This will come up again in Chapter 5.) Adopting this

refined understanding of material correspondence, were we to endorse the materiality thesis, we should hope to be able to at least roughly rank research programs in terms of their achieved degree of material object-target correspondence. However, this gets difficult if we try to think through our two examples of inferences from the Lenski system in these terms.

In the high mutation rates case, it is not so difficult. The inference is from experimental populations of *E. coli* to natural populations of *E. coli* and *Salmonella*: The mechanism posited for explaining the evolution of high mutation rates in the former might be the same mechanism responsible for the evolution of high mutation rates in the latter. There is straightforward material correspondence, in the sense outlined above, between object and target: *E. coli* in the laboratory belong to the same type as *E. coli* in nature, at a fine-grained level of classification of biological types: species. *E. coli* and *Salmonella* belong to the same type at a level which is not as fine-grained, but still, arguably, significantly fine-grained.[25] Thus, in the high mutation rates case, the object (experimental populations) and target (natural populations) in question differ in their exact genetic makeup and environments, but they correspond materially according to the scheme laid out above for making sense of the notion of material correspondence.

In the punctuated evolution case there is not such straightforward material correspondence. It is not even clear how to go about evaluating it. We know what the object is: The same set of twelve experimental populations of *E. coli* as before. But what exactly is the target? How is it classified materially? In this case, the claim was that rare, beneficial mutations sweeping to fixation explain the punctuated evolutionary dynamic in the laboratory populations, and that this same process could explain punctuated

[25] While phylogenetics can get messy for bacteria, *E. coli* and *Salmonella* are in any case closely related; see, for example (Fukushima, Kakinuma, & Kawaguchi 2002).

equilibrium in nature. This claim rests on the experimental populations sharing properties of their evolutionary history with an arbitrary set of populations, traces of whose evolutionary history are left in the fossil record. Those properties, in particular, were (1) displaying a certain macroevolutionary trend, (2) existing in a constant environment, and (3) having rare beneficial mutations. Note that none of these properties has to do with what the population is "made of:" its phylogenetic classification, or any specifics of its phenotype. How are we to assess material object-target correspondence here? Both object and target comprise or once comprised living organisms; that is a start. But the target is not identified in such a way that its material correspondence to *E. coli* can be straightforwardly evaluated.

Material object–target correspondence is not necessarily a characteristic feature of experiments, either in design or in retrospective analysis. This case illustrates that nicely because it can be said of the exact same experimental object that it corresponds materially to one target, but it is unclear whether it corresponds materially to another, or how to even evaluate such correspondence.

A separate issue, however, is: What conditions would need to hold for the inferences in question, from the laboratory population of *E. coli* to these different targets in the natural world, to be valid? In the high mutation rates case, the inference relies at least in part on the object corresponding materially to the target, in the sense outlined above. It relies on mutator hitchhiking actually being the mechanism responsible for the evolution of high mutation rates observed in the laboratory populations, and on the reasonability of assuming that that same mechanism could explain the same observation in closely related natural populations. Here the researchers seem to rely on material correspondence in the way Morgan and Guala had in mind: Identifying a mechanism in a

physical object of study (experimental populations) allows you to make an inference about materially corresponding targets in the natural world (natural populations in the same biological/phylogenetic class: asexual pathogenic microbes).

In the punctuated evolution case material correspondence is not playing such a role. There the focus is on evolutionary mechanisms and environmental particulars, not what the target populations are "made of" in anything like the material correspondence sense. For the punctuated evolution inference to be valid, the population-level mechanisms and environmental conditions responsible for the evolutionary dynamics in the laboratory populations would need to correspond to mechanisms and conditions responsible for evolutionary dynamics in the relevant natural populations. But physical, physiological, or phylogenetic particulars of the populations in question are not figuring in to the researchers' reasons to think that is the case.

I agree with Parker (2009) that material object–target correspondence does not necessarily entail greater epistemic value. She discusses cases like climate modeling to make this point. In these cases, we make predictions about climate change which are as reliable as possible by building complicated large-scale computer models; we cannot generate similar predictions by experimenting on the actual global climate, and there is no reason to think that creating "same stuff" laboratory analogues would do anywhere near as well as the computer simulations. Material correspondence is one (often good) route to grounding valid inferences from objects to targets. But it is not the only route; material object–target correspondence is neither necessary nor sufficient for valid scientific inferences.

My discussion above establishes a further point beyond Parker's: It does not even follow from the fact that we have a material system as our object of study that material

correspondence is doing, or is even meant to be doing, the work in validating an inference. If we want to assess inferential power in cases like the punctuated evolution one, we need to look somewhere other than success or failure at achieving material correspondence.

The picture is further complicated because proponents of the materiality thesis compare experiment to simulation in the broader sense (Guala 2002; Morgan 2005), while its opponents have sometimes followed suit (Winsberg 2010) and sometimes focused only on computer simulation (Parker 2009). I mentioned in Section 4.1 that I do not think there is a clear distinction to be drawn between experiment and simulation in the broader sense. The cases under discussion do a nice job of showing why. If we consider the difference between the two examples from the Lenski experiment from the perspective of the materiality thesis, we can generate some puzzling questions. Should we think of both cases as experiments because their objects are the same kind of physical experimental system in a laboratory (we call the research area "experimental evolution," after all)? Should we think of the high mutation rates case as an experiment and the punctuated evolution case as a broad-sense (physical) simulation? The rationale for this might be that the former takes the object as an instance of some target and makes inferences grounded in material (and, in this case, phylogenetic) similarities, while the latter arguably takes the object as more of a representation or model, and makes inferences grounded in more formal similarities. Or should we think of both cases as broad-sense simulations, because the object in question is a population of model organisms, a kind of concrete theoretical model?[26]

[26] Some (Frigg and Hartmann 2012; Harré 2003; Humphreys 2004; Weisberg 2013) say so of model organisms, in any case. Levy and Currie (forthcoming) argue that model organisms play a different role in scientific inferences than traditional concrete theoretical models.

Plausible arguments could be made for affirmative answers to all three of those questions, and people argue about just these kinds of questions; see, for example, discussion in Guala (2002), and the references mentioned in footnote 26. But if we are interested in accounting for how and why cases of scientific inquiry differ in inferential power, these sorts of questions are red herrings. This discussion of laboratory natural selection experiments, and the different kinds of inferences that can be made from the same object of study, highlight the fuzziness of the distinction between experiment and broad-sense simulation. Whether we classify a case of scientific inquiry as an experiment, a simulation, a hybrid, or explicitly both at once, per se, should not make a difference to how we judge its inferential power. If one of the inferences in question is more licensed than the other, the right way to think about this is in terms of specific details of the case and how well the object captures features relevant to making a good inference about the target in question—not in terms of how we categorize it.

It does not follow from a case of inquiry being categorized as an experiment that any inference it makes is (actually or intended to be) based in its degree of material object-target correspondence. Furthermore, categorization as an experiment or a simulation alone does not tell us anything about inferential power. We should not look to the experiment/simulation distinction as a basis for in-principle judgments about inferential power. There are certainly particular contexts in which experiments put us in a better position for making valid inferences than simulations. But it does not follow that experiments are always better generators of inferential power in principle.

Again, the discussion in the preceding paragraphs was not supposed to be an assessment of the validity of the inferences in the punctuated evolution and high mutation rates cases. The question was: If we adopt the material correspondence view about what

matters for grounding external validity, what can we say about whether or not the experimental and target systems are made of the same stuff? And the answer is: First, it is not obvious what this even means when we get down into the details, and second, it looks like we should be concerned with more than just material correspondence. The material correspondence (designed or not) of experimental systems to their targets is not alone a good grounds for assessing the external validity of experiments (versus simulations, or period). I say it is not *alone* a good grounds because sometimes material correspondence does matter; sometimes it is crucial. If we know a little bit, but not much, about some target of inquiry in the world, often a great way to learn about it is to intervene on it directly, or set up something very much like it in the lab that we can intervene on. But material correspondence matters in these cases not because it is material correspondence per se, but because this sometimes happens to be the best route to capturing the similarities most relevant to grounding the inference one wants to make.

4.4. Surprise

Even if we agree on the points I have made so far in response to the materiality thesis, people still want to say that there is a further difference between experiments and computer simulations which affects their epistemic value: Simulations cannot surprise us the way experiments can (from now on I will refer to this as *the surprise claim*). The thinking behind the surprise claim has to do, again, with the nature of experimental objects of study: While experimenters usually design at least some of their object's parts and properties, they never design all of them, and in some cases they design none of them, such as in some field experiments. A simulationist's object of study, on the other hand, is a

model: She made or programmed it herself, so knows all of the relevant facts about its parts and properties. It is thought that experiments, in virtue of these facts about the nature of their objects, are a superior source of surprises compared to simulations.

Surprise plays an important role in what we value about science. People have talked about surprise in a number of different ways. There is a term used in cognitive neuroscience, 'surprisingness,' which defines that as a function of the absolute value of the difference between observed and expected outcomes (Hayden et al. 2011). Another way to think about surprise is an epistemic change by which we believe that some outcome is improbable or are indifferent to whether or not it might obtain, and then realize that it is in fact probable, or it is in fact the case. There are also cases of surprising results in science which seem to go beyond changes in our beliefs about probabilities; for example, cases of entirely new things being added to our ontology.

Those are some existing ways to think about surprise in science. I do not want to hang too much on a particular definition of surprise. What I have in mind broadly in this discussion is the sort of novel or unexpected result, behavior, discovery, or insight regarding scientific objects of study, characteristic of what's really interesting about doing science. Surprising results are either contrary to what we expected, or introduce possibilities we didn't even know were on the table.

A common claim about a difference in epistemic value between experiments, compared to simulations, is the surprise claim, which says that as a matter of principle experiments are a superior source of surprises. Not many people have put this claim in writing, but it comes up all the time in discussions of the difference between experiments and simulations. The strongest form of the surprise claim would hold that simulations

cannot genuinely surprise us at all. More commonly, the claim is that simulations and experiments differ, either qualitatively or quantitatively, in their capacity to surprise us.

Paul Sniegowski (personal communication, cited with permission), discussing the difference between experiment and simulation in evolutionary biology, writes: "Although surprises do emerge in simulations, in general what goes into a simulation is well known and surprises are not anticipated. In contrast, surprises and exceptions to anticipated results are fairly common in experimental systems." Morgan endorses another version of the surprise claim, arguing that while simulations may be able to surprise us, experiments can both surprise and confound. She writes:

> [N]ew behaviour patterns, ones that surprise and at first confound the profession, are only possible if experimental subjects are given the freedom to behave other than expected. […] This potential for laboratory experiments to surprise and confound contrasts with the potential for mathematical model experiments only to surprise.[27] In mathematical model construction, the economist knows the resources that went into the model. Using the model may reveal some surprising, and perhaps unexpected, aspects of the model behaviour. Indeed, the point of using the model is to reveal its implications, test its limits and so forth. But in principle, the constraints on the model's behaviour are set, however opaque they may be, by the economist who built the model so that however unexpected the model outcomes, they can be traced back to, and re-explained in terms of, the model. (2005, pp. 324–5)

Note that Sniegowski is making what I called a more quantitative version of the surprise claim. He is saying that surprises are commonplace in experiments and rare in simulations. Morgan, on the other hand, is making a more qualitative version of the claim. By "confounding" she has in mind motivating a researcher to seriously question her relevant background theoretical knowledge, as opposed to merely seeing something she was not quite expecting to see.

---

[27] By "mathematical model experiment" Morgan means what I am calling a simulation: a study of a model (in this case, a mathematical or computational model) with some dynamic temporal element. Her focus here is on examples from experimental economics, but she intends her discussion in the paper to apply to experiments versus simulations in general.

I take the idea behind the surprise claim, in its various versions, to be generally based on the following line of thinking: The objects of study in simulations are computational or mathematical models, while the objects of study in experiments are physical systems in the laboratory or the field. While experimenters usually design at least some of the parts and properties of their objects of study, they never design all of them, and in some cases they design none of them (for example, in some field experiments). So details of the object of study come along for free, for the experimenter, which she did not knowingly put there herself. A simulationist, on the other hand, has a different relationship with her object of study (the model): She made or programmed it herself, so —the thinking behind the surprise claim often goes—she knows all of the relevant facts about its parts and properties. It is thought that experiments, in virtue of these points about their objects of study, are thus superior sources of surprise as a matter of principle: either simulations lead to surprises rarely (or never) in comparison, or experiments can surprise us in ways that simulations cannot. That is what I take to be the core intuition behind the surprise claim. An extreme version of the surprise claim, though I do not think that anyone actually endorses this view anymore, would be that whenever you do a simulation you are just learning things you already knew.

4.4.1. Response to the Surprise Claim

People making the surprise claim have in mind some quantitative or qualitative difference between experiments and simulations, regarding to their capacity to surprise us. I want to get more precise about what we are talking about here, and shift the discussion to thinking in more detail about potential kinds of sources of surprise in scientific objects of study.

I will focus on two kinds of sources of empirical surprise. The first is *unexpected behavior*: surprising states or phenomena exhibited by a scientific object of study over the course of studying it. The second is *hidden features*: sources of surprise that in some important sense can be said to have "been there all along." These are features of the object of study itself, which a researcher was genuinely unaware of prior to studying it. Unexpected behaviors and hidden features are not mutually exclusive categories. The former is a source of surprise in its own right, and it is also sometimes (but not always) a sign of the latter. Unexpected behaviors in an object of study surprise us, for example, by displaying some state of that object which we did not anticipate it would be in, or by failing to straightforwardly refute a hypothesis in the sort of way we were expecting. Sometimes unexpected behaviors can also motivate us to dig down and question our knowledge of the fundamental workings of the object of study itself. When this happens, we might learn that those unexpected behaviors were actually just caused by artifacts or bugs. But sometimes, when we investigate them further, we will uncover important hidden features of our object of study which we genuinely did not know about before we began studying it. These can include mechanisms, causal factors, properties, key components, or variables.

Unexpected behaviors are found in experiments all the time. Lenski and colleagues' work is again a great source of examples. A number of the many publications from the Lenski experiment are based on unexpected behaviors the populations have exhibited over their 27+ years (60,000+ generations) of evolution. Two noteworthy examples are the evolution of surprisingly high mutation rates, two orders of magnitude higher than the ancestor's, after 10,000 generations, and the evolution of citrate utilization after 31,500 generations (these were discussed in Sections 2.1.1 and 2.1.3).

Another kind of example of unexpected behavior comes from a more recent laboratory natural selection study of the evolution of mutation rates. Gentile and colleagues (2011) looked at the relationship between mutation rates and fitness evolution in engineered "single mutator" and "double mutator" genotypes of *E. coli*, which have respectively high and extremely high mutation rates compared to the wild type (Figure 5).

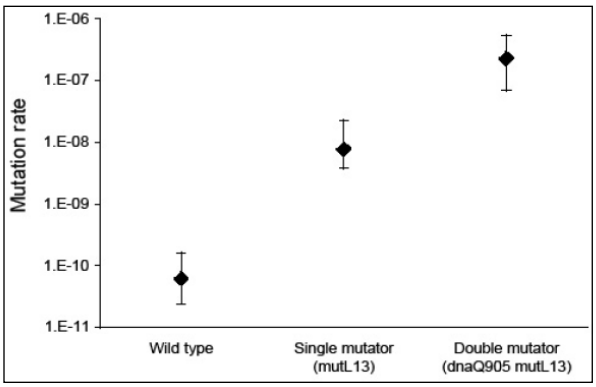

Figure 5: Mutation rates of wild type, single and double mutator *E. coli* (from (Gentile 2012); see also (Gentile et al. 2011)). Estimates are of the per base pair, per generation genomic mutation rates with 95% confidence intervals for wild type *E. coli*, single mutators with the *mutL13* allele (which confers deficiency in mismatch repair), and double mutators with *mutL13* and *dnaQ905* alleles (the latter confers deficiency in DNA proofreading). Single mutators have a genomic mutation rate 100–fold higher than the wild type; double mutators have a genomic mutation rate 45–fold higher than single mutators.

One would think that a population with a genomic mutation rate as high as the double mutator's, which is 4,500-fold higher than the wild type, would not last long. Thinking about mutation rates intuitively, there are plenty more ways to mess something up at random than ways to improve it. Populations in nature, as far as we know, never have genomic mutation rates this high, and theories such as the error catastrophe hypothesis (Eigen 1971, 2002) and Muller's ratchet (Muller 1964) predict that populations with very high mutation rates will decline in fitness and go extinct. But surprisingly, using a serial-transfer protocol similar to that in the Lenski experiment, the double mutators remained viable for over 2,500 generations (Gentile 2012; Gentile et al. 2011), and stopped then only

because the experimenters stopped propagating them. This case is worth mentioning, in addition to the cases from the Lenski system mentioned above, because they are examples of different kinds of unexpected behaviors: The evolution of high mutation rates and citrate utilization were cases of unexpected features arising over time in the object of study, while this is a case of things going explicitly otherwise than what we might have thought from the getgo given the relevant background theory.

Unexpected behaviors also occur all the time in simulations. Examples abound in the area of agent-based modeling. In agent-based models, also known as individual-based models, individual agents and their properties are represented and the consequences of their dynamics and interactions are studied via computational simulation. Common applications of agent-based models include in ecology and the social sciences, where agents can represent individual organisms and their interactions, locations, behaviors, life history traits, and so forth. Behavioral patterns can emerge from simple initial conditions comprising agents, their properties, and their interactions, such as complex cycles of fluctuation in population size or flocking behavior (Epstein & Axtell 1996; Grimm & Railsback 2013; Railsback & Grimm 2011).

One example of unexpected behavior from simulations, keeping with the theme of studying evolving populations, is evolved predator avoidance in Avida. Avida is an agent-based model in which self-replicating "digital organisms" compete for resources in the form of computer memory (Ofria & Wilke 2004). Ofria and colleagues describe a case in which they wanted to study a population that could not adapt, but would accumulate deleterious or neutral mutations through genetic drift. Agent-based models are idea for this kind of study: Researchers can examine each new mutation as it occurs by running a copy of the mutant agent in a test environment and measuring its fitness. The test allowed

them to identify agents in the primary population with beneficial mutations and kill them off, which would in theory stop all future adaptation. Surprisingly, however, the population continued to evolve. It turned out that the agents had developed a method of detecting the inputs provided in the test environments, and once they determined that they were in a test environment, they downgraded their performance. As the authors put it, the agents in the model "evolved predator avoidance," the performance downgrade being an adaptation to avoid being killed.[28]

It makes sense that simulations and experiments can both involve sources of surprise in the form of unexpected behaviors, if we consider their methodological points in common. An experiment starts with choosing or designing an object of study and specifying a protocol. A simulation starts with the object of study, a model, in some initial state with a set of transition rules specifying how it will update to future states. In both cases, a researcher sees what happens to her object of study over time. The examples of unexpected behaviors I just discussed were all cases of subsequent states or properties of the object of study differing in surprising or unexpected ways from its initial states or properties.

The extreme version of the surprise claim—that a researcher cannot be genuinely surprised by her simulations because she programmed them, so knows everything about them—is plainly false. A simulationist will often, but not always, know everything about her model's initial conditions and transition rules. A straightforward case in which she might not know everything is when she did not write the model herself, so is ignorant of aspects of how it was programmed or how it works. But there are more interesting reasons why she might fail to know everything. For example, she might be writing the model in a

[28] One could argue about whether this is the most plausible way to describe what went on here, but that is beside the point; in any case this is a clear example of unexpected behavior in a simulation.

high-level programming language and fail to understand all of its low-level details. Or she might program the model in a way that leads to its initial conditions having unintended features, or its transition rules entailing unintended consequences. Furthermore, very complex models are often written by teams rather than individual (for example, in climate modeling); in some such cases, no individual researcher might be said to understand everything about the model's initial conditions and transition rules.

In any case, knowing "everything" about a model's initial conditions and transition rules does not entail knowledge of its future states. Setting an initial state and deciding which rules will govern its change over time does not tell you what will happen—that is why we must run the simulation. Similarly, finding out as much as you can about an experimental object of study and sorting out all the details of your protocol does not tell you what will happen in the experiment. Both experiments and simulations can exhibit unexpected behaviors. Any study of a system with an initial state and subsequent states has at least the potential to surprise us, because it contains potential sources of unexpected behavior as it changes (or fails to change) over time.

I now turn to hidden features. Unlike unexpected behavior which an object of study manifests over the course of studying it, these are features an object of study already had, in some sense, which a researcher was genuinely unaware of when she embarked on a study. Hidden features are always accompanied by unexpected behaviors, but the converse is not true. Hidden features are discovered as a result of investigating unexpected behaviors, but investigating unexpected behaviors does not always lead to discovering hidden features.

A perfect example of surprise in the form of a hidden feature is the discovery of transposable genetic elements. Barbara McClintock, over the course of her studies of the

genetic basis of maize patterns, discovered that the gene regulating the maize's mottled pattern also made its chromosomes break. In the process of examining this breakage, she eventually discovered that genes can move from one place to another on the chromosome, with a sort of cut and paste mechanism, refuting the earlier belief that genes' positions on the chromosome are fixed (McClintock 1951). McClintock discovered transposable elements over the course of her studies of maize plants, but in an important sense she was discovering a hidden feature of the genome that had been there all along, which she didn't know was there; nobody knew it was there.

Simulations can also contain hidden features. Here is an example: The agent-based model Sugarscape is a simple model consisting of cells in a grid. Every cell can contain different amounts of sugar or spice (resources), and there are agents (red dots) which can move around the grid. The basic setup of the model is that with each time step, agents look around for the nearest cell in their neighborhood with the most sugar, move, and metabolize. These simple local rules can give rise to population-level features that look remarkably like the macrostructures we see in societies of living organisms: structured group-level movement, carrying capacities, distributions of wealth, migration patterns, and so forth. The model's creators discuss these results as follows:

> Now, upon first exposure to these familiar social, or macroscopic structures… some people say, "Yes, that looks familiar. But I've seen it before. What's the surprise?" The surprise consists precisely in the emergence of familiar macrostructures *from the bottom up*—from the simple local rules that outwardly appear quite remote from the social, or collective, phenomena they generate. In short, *it is not the emergent object per se that is surprising, but the generative sufficiency of the simple local rules*. (Epstein & Axtell 1996, pp. 51–2)

Now, one might think: That's not a hidden feature. You had to run the model to see the macrostructures, they were not just sitting there in the initial conditions. That is true, but there is something revealing in what Epstein and Axtell say here, in the italicized last

sentence: The surprise is not so much in the details of the behavior itself, but in the fact that these simple local rules are sufficient to generate it. This object of study which looks very simple has generative properties that one would have never known about until studying it. And the interesting lessons in this case come from studying that fact and how it works, not the "familiar macrostructures," per se.

Another example supporting the idea of hidden features in a simulation comes from Conway's Game of Life. The Game of Life is a cellular automaton, a simple model consisting of a collection of cells on a grid which evolve in discrete time steps, according to rules based on the states of their neighboring cells. Cellular automata have been studied since the 1950s; they were originally thought of as possible representations of biological systems, and went on to be used to study a wide range of issues in computation and complexity science. The Game of Life is a two-dimensional grid whose cells can be in one of two states: "on/living" or "off/dead" (Figure 6a). The update rule is simple: an "on" cell will remain on at next time step only if exactly two or three neighbors in its Moore neighborhood (the eight cells immediately adjacent and diagonal) are on; otherwise it will turn off. An "off" cell will turn on only if exactly three neighbors in its Moore neighborhood are on.

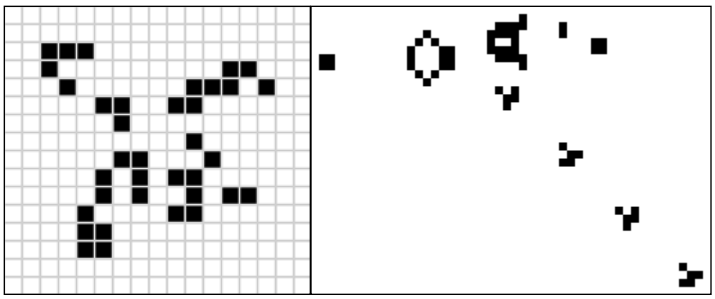

Figure 6: (a) A representative grid of cells in the Game of Life. (b) A glider gun, with a stream of gliders moving off toward the lower right.

Understanding the two possibilities for cell states, and knowing the transition rules, in one sense tells you everything you need to know about how the model works. But once the simulation begins, it produces surprisingly complex results. The simple rules and initial conditions generate an amazing number of different patterns, with organized structures and entities apparently persisting at a level higher than the individual cells (Bedau 2008; Dennett 1991; Weisberg 2013). A number of surprising results have come from studying the Game of Life. John Conway, the model's creator, did not think that the model was capable of producing an infinite number of cells, and offered a fifty-dollar prize in 1970 to whomever could prove him wrong (Weisstein 2013). He was proven wrong by the discovery that certain initial conditions give rise to "glider guns," configurations of cells that spit out stable patterns, called gliders, which move off into infinity through the two-dimensional grid, maintaining their structure as they go (Figure 6b). The ability to produce an infinite number of cells from a finite number of initial "on" cells is an unexpected behavior of the Game of Life. The glider guns can be thought of as a hidden feature, at the macrostructure, in the model.

This difference between unexpected behaviors and hidden features is another way to articulate the kind of idea I take it Morgan had in mind regarding the difference between surprise and confoundment. Namely, there are plenty of situations in which we can be surprised by unexpected behaviors, but only in special circumstances do surprising results cause us to dig down and question our knowledge of the workings of the object of study itself, or learn something about it that we genuinely did not know going into the research program. Unexpected behaviors, in the sense discussed here, are the kinds of things that surprise, in Morgan's sense. The process of investigating particular unexpected behaviors, searching for possible hidden features, in the sense discussed here, would seem

to correspond roughly to Morgan's notion of confoundment. So, I think Morgan's distinction between surprise and confoundment is important, and I'd like to think that I have preserved some version of it in my account. However—and this is the key point—unlike Morgan, I am arguing that neither form of surprise is unique to experiments.

It does seem, though, that studies of material systems arguably put us in a better position to uncover a particular kind of hidden feature. Here is what I have in mind. We can talk about hidden features existing in scientific objects of study at different levels. In particular, we can distinguish among the micro-level and the macro-level. I am keeping the terms as general as possible here because, depending on the area of inquiry, exactly how we think of these levels of organization in our object of study, and exactly what we call them, will vary. The micro-level might be thought of as the level of individuals, molecules, or atoms. The macro-level might be thought of as the level of aggregates, populations, or wholes. In population genetics the relevant levels might range from allelic to population; in ecology, from individual to community; in chemistry, from atomic to aggregate; and so forth. I do not intend to give any particular metaphysical weight to the distinction among these levels, but just to point out that we can think of most scientific objects of study in these sorts of terms (it need not be an objective matter of fact exactly what the relevant levels are in a given research area). Thinking in terms of different levels of organization in objects of study matters, for our purposes here, because there is a sense in which micro-level hidden features seem (1) more hidden than macro-level ones, and (2) potentially unique to experiments.

The transposable elements case involved the discovery of a micro-level hidden feature, the mechanism of genetic element transposition. The examples of generated macrostructures in Sugarscape and glider guns in the Game of Life involve macro-level

hidden features. It seems tempting to say that here might be the grain of truth in the surprise claim: There is this particular kind of source of surprise that might be unique to experiments, namely, these micro-level hidden mechanisms. People think there is something special about studying physical systems; I mentioned that at the outset. Part of the intuition there is that a lot of stuff comes along for free when you adopt a physical system plucked from the world, as opposed to a computer model written from scratch, as your object of study. People usually make this point in the context of arguing for why experiments put us in a better position to make inferences about the world (see my earlier discussion of the materiality thesis). But it also bears on the surprise claim.

However, there are plausible counterexamples to this idea that surprises in the form of micro–level hidden features come only from experiments. Here is an example from nanoscale physics. Lenhard (2006) discusses a molecular dynamics simulation which uncovered properties of gold nobody previously knew about. In particular, when nickel tips are held against gold plates and slowly removed, the gold deforms to make nanoscale wires of gold atoms (Landman et al. 1990). Lenhard quotes an interview with the model's creators: "That gold would deform in this manner amazed us, because gold is not supposed to do this" (Lenhard 2006, p. 606). The simulation results were confirmed later by atomic force microscopy. Lenhard uses this example to argue for the point that simulations, like experiments, can be "epistemically opaque," even when the person running the model of study built it themselves "from scratch." So this is a counterexample to the idea that micro-level hidden features can be uncovered only in experiments. Though it seems plausible that discovery of micro-level hidden features in simulations might be particular to cases like this, where the model is based on a significant well-

known body of theory about the physical microstructure of the target of inquiry in question.

The upshot of all of this discussion is that the surprise claim is false as a generalization. Experiments and simulations both have the potential to give rise to unexpected behaviors. While in particular contexts there might be reasons to think that experiments will lead to more unexpected behaviors than simulations (or the converse!), there are not grounds for claiming that this is the case in general. I have given reasons to think that both experiments and simulations can lead to the discovery of hidden features. Though on this latter point, it still seems right to say that simulations do not contain sources of a particular kind of surprise—namely, micro–level hidden features—as often as experiments do.

Within the contexts of certain research areas, experiments may well be a superior source of surprise over simulations. This might be the case especially in new fields where exploratory experiments are the (current) chief means of learning about the world, good simulations would have to be data-driven, and we do not have the data yet. This is just to underline that the target of my objection is people who endorse the surprise claim as generally true across science, and who take it to support claims for the general epistemic superiority of experiments.

4.5. Experimenters Almost Never Study Their Targets Directly

There is a view which is in the background in arguments for the materiality thesis and the surprise claim, but which one could hold without endorsing either. This is the relatively common view that a key difference between experiments and simulations is that experimenters study their targets directly, while simulationists do not. Winsberg (2009, p.

577) quotes Gilbert and Troitzsch as holding this view: "The major difference is that while in an experiment, one is controlling the actual object of interest (for example, in a chemistry experiment, the chemicals under investigation), in a simulation one is experimenting with a model rather than the phenomenon itself."

This kind of view underlies the claim that reproduction versus representation characterize object–target relationships in experiments versus simulations, respectively, discussed in Section 4.2 above. It also comes up for proponents of the surprise claim, and could be a line of response to my response to the surprise claim in Section 4.4.1. The idea would be that when you discover a hidden feature in an experimental object of study, you are learning something about your target. But when you discover a hidden feature in a simulation you are learning something only about your model. For that reason, the objection goes, experiments still put us in an epistemically privileged position with respect to their capacity to surprise us. This objection is not necessarily about *how often* experiments versus simulations lead to surprises, or how often they lead to the discovery of genuine hidden features. The objection regards the epistemic payoff of those hidden features. That is, it says that the productivity of surprises from experiments is worth more because it tells us about targets in the world. The claim is that when a researcher discovers a hidden feature in a simulation, the burden is on him to show why it tells us anything about the world outside the model.

This view is wrong. Well, it is not entirely wrong: It is true that simulationists do not study their targets directly. They study models, which stand in for their targets. But experimenters almost never study their targets directly, either. In physics and chemistry, objects of study in the laboratory are often instances of the target entities or phenomena in the natural world, like particular subatomic particles or elements or kinds of reactions.

But even then experimenters need to do some work to show why the conditions that apply to those entities or phenomena in the laboratory apply to them in general out in nature as well, or (sometimes) whether they are even identifying the correct entities or phenomena in the laboratory (on this latter point see related discussion in Galison 1987). In biology and biomedical research, there is often even more work to be done to show why inferences from the object of study to a target of inquiry in the natural world are licensed, as when using *Drosophila* to study genetics in general, or a small group of clinical trial volunteers to study potential future drug-takers in general. When computer models are the objects of study, scientists always have to do the work to show why inferences to the world outside the object of study are valid. When experimental systems are the objects of study, they almost always do too.

Only in very rare cases can experimenters be said to study their targets directly. The only sorts of cases where this applies are cases where (1) the target is delineated in particular rather than general terms and consists of a small, clearly delimited set of entities, and (2) the experimenter is studying exactly that set of entities as her object. Cases where this might hold include studies of very small populations, for example, certain field studies in anthropology, where the goal is to say something about a small, clearly delimited population of humans and researchers engage with every member of that population as their object of study; or biomedical studies of the only 50 people in the world with an extremely rare genetic disorder for the purposes of making inferences about people with that genetic disorder.[29] Other rare cases that might meet these conditions include chemistry experiments in which researchers create a new synthetic element in the laboratory, for the purpose of making an inference about how that element behaves, and

[29] Even here, though, there could be problematic issues involved in making inferences about past or future people with the disorder in question.

the only existing instance of that element is the one which they are studying. For example, in 2014 researchers claimed to have created element 117 (ununseptium) (Khuyagbaatar et al. 2014); it is very difficult to create this synthetic element in the laboratory and the element exists for only a fraction of a second before falling apart. But people created it and wrote a paper about it. In these sorts of cases, where the only proper instance of the target in the universe (as far as we know) is exhausted by the experimenter's object of study, it is correct to say that the experimenter is studying her target directly. In the vast majority of experiments, this is incorrect.

These sorts of cases are rare exceptions, and not at all the norm. In all other cases of scientific research, the object of study is standing in for the target of inquiry. There are many ways for one thing to stand in for or represent something else, and it is beyond the scope of this discussion to enumerate all these various ways and exactly how to understand them (but see Frigg 2006, Van Fraassen 2008 and others). The key point here is just that a scientist always has to do some extrapolation work to show why her object is an appropriate stand-in for her target. This applies whether her object is a physical system in a laboratory, a population in the field, or a model visualized on a computer screen. This applies whether the difference between her target and object is that one is a biological population in nature and the other is computer code, or that one is a biological population in nature and the other is a biological population in a test tube. If we were to define experiments as cases where researchers study their targets directly, then very few cases of scientific research would count as genuine experiments.

It is worth mentioning another exception to the view that experimenters study their targets directly, simulationists do not: There are also rare cases where simulationists *do* study their targets directly. For example, it can be the case that a simulationist's

ultimate target of inquiry is not some system in the natural world, but the model under study itself. Weisberg calls these cases "targetless models," where "[t]he only object of study is the model itself, without regard to what it tells us about any specific real-world system" (2013, p. 129). The Game of Life (see discussion in Section 4.4.1) is a perfect example. The model is not meant to represent some particular target system in the natural world. Rather, it is studied as an interesting case in its own right of properties of interest to researchers in artificial life and computer science, like emergent dynamics and universal computation.

4.6. What's the Difference between Experiment and Simulation?

To underline a point which should be clear by now: When we are talking about broad-sense simulation and comparing experiments to physical simulations (studies of physical models), there is no interesting or important difference between experiment and simulation. I have discussed cases which we might think of as either experiments or broad sense simulations. It does not ultimately matter, for the purpose of evaluating the inferences in question, whether we call these experiments or simulations. The same is true for any other study where researchers are intervening in a physical system (their object) for the purpose of making some inference about some other physical system (their target). Other than in very rare cases like the ones mentioned in Section 4.5, scientists' objects of study are always stand-ins for their target. They almost never study their targets directly. The important question for evaluating what we can learn from a given study of a physical system is not "Is it a simulation or an experiment?" The important question is, how is this object of study being used to generate or justify inferences about the target of inquiry in

question. I agree with Winsberg (2009, p. 591) when he says that "[h]ow trustworthy or reliable an experiment or simulation is depends on the quality of the background knowledge, and the skill with which it is put to use, and not on which kind it belongs to." There is no epistemic or methodological distinction between experiments and physical simulations. Focusing on which kind a case of scientific inquiry belongs to focuses us on the wrong issues.

That was all about experiments versus physical simulations. When we're talking about experiments versus computer simulations rather than broad-sense simulations, there is an important methodological difference at play, namely, the difference between studying a physical system and studying a computer simulation. This matters for pragmatic reasons. Most often, doing an experiment will be more costly than doing a simulation. The supplies, reagents, and person hours needed to run a laboratory experiment tend to cost significantly more than running a model on a computer. (I say tend to because even here there are exceptions: running most middle-school chemistry experiments costs significantly less than running meteorologists' climate models). Simulations can allow one to observe an object of study's dynamics over time much more quickly than doing so in a real-time experimental system.

This pragmatic advantage can come with epistemic costs. Many people have the intuition that it always comes with epistemic costs; this is an important part of the intuition which the materiality thesis tries to explain: The idea is that studying a model as opposed to a material system involves sacrificing realism, and sacrificing realism reduces epistemic value. Again, this is a good intuition in contexts where we know relatively little about the features of our target of inquiry relevant to designing a good experimental system or model. But science is not always operating in such contexts. One example of a

context in which there is no such epistemic cost associated with simulation is the study of molecular bond angles in chemistry. We know enough about chemical bonding to answer questions via mathematical modeling and computer simulation about how atomic substitutions will affect the bond angle in a given molecule; for example, swapping atoms of phosphorus for atoms of arsenic in the molecular backbone of DNA (as in Denning & MacKerell 2011). For answering questions about straightforward atomic substitutions in familiar molecules, we would not be in a better epistemic situation were we to carry out the relevant experimental manipulation (and it would certainly be far more pragmatically costly). So again, the point about the epistemic costs of simulation is a point that holds in many contexts. But it is not an in-principle epistemic difference between experiment and simulation.

The methodological difference between experiment and simulation is not purely pragmatic. It matters for making judgments about epistemic value—but only in a context-sensitive way. All of science is about engaging with some object of study to learn about some target of inquiry, and very rarely are the object and target identical. We should not look to the experiment/simulation distinction alone to tell us anything in principle about the epistemic value of cases of scientific inquiry.

A final point about the difference between studying experimental systems and computer models: Even regarding this version of the experiment/simulation distinction, people have been hasty in drawing sharp methodological lines (and using these, in turn, as a basis for conclusions about the epistemic superiority of experiments). While it was once common for individual scientists, laboratories, or even entire subfields to focus on one or the other, experimental and computational methods are now increasingly combined. While of course there is the methodological distinction I just mentioned, the

identity of the object of study, the methodological overlap between experiment and computer simulation in another sense is often significant. The views on experiment versus simulation cited above often talk as if researchers choose to do one or the other, as their main program of research or even within the context of a given study. But this is increasingly not the case. Computer simulations are often a key part of the experimental process. Of course, one way computation is used in biology is to solve mathematical models, like population- and Mendelian-based models of evolutionary genetics. But simulations can play plenty of other roles in the experimental process as well. For example, they are used as initial steps in LNS experiments to help figure out which variables to fix (Roff & Fairbarin 2009), or to explore theoretical questions about spatial structure in experimental microbial communities in tandem with studying those communities themselves (Kerr et al. 2002).

I have argued that we should not look to the experiment/simulation distinction to tell us anything in principle about epistemic value. I have shows that two senses in which experiments are commonly thought to have epistemic privilege over simulations—inferential power and capacity to generate surprises—do not generalize across science. Studying a material system as opposed to a computer model does not automatically entail better inferences. In Chapter 5 I will begin to develop an account of where we should look instead.

5. Conclusion: Evaluating Inferences about the Natural World

The views I responded to in Chapter 4 held that we can look to the experiment/simulation distinction to tell us something in principle about the validity of scientific inferences. I argued against the claim that experiments are superior to simulations as a generalization across science, both in the sense of having greater inferential power and being a superior source of surprises, and furthermore that we should not rely on the experiment/simulation distinction, in and of itself, as a basis for making judgments about the epistemic value of cases of scientific inquiry. The trend toward carving up scientific methodologies into clear categories, and using those categories as bases for such judgments, focuses on the wrong issues.

In this concluding chapter I begin to sketch an account of where we should focus instead when evaluating inferences. The starting point, rather than the methodological category of the case in question (was it an experiment, a simulation, an observation…?), should be the extent to which relevant similarities have been captured between the object and target, along with context-sensitive information about background knowledge and the scientific question at hand. Methodological considerations like degrees of control, intervention, material object-target correspondence, and the identity of the object of study (is it an experimental system on a lab bench, a computer model, a population in the field…?) can certainly be focal points in these evaluations. But these methodological aspects should play a context-sensitive role. The conclusions we draw from them depend on the question at hand and the researcher's epistemic relationship to her object and target. This framework allows us to understand how doing an experiment or a simulation

can have strong advantages—but not purely in virtue of being an experiment or simulation, per se. I will discuss three sorts of key considerations: relevant similarities, how much we know (about the target and the object), and the importance of realism versus control.[30]

The core focus of evaluating inferences without problematically focusing on methodological categorizations should be to look at the extent to which a given object of study is relevantly similar to the target of inquiry in question. In recent work Parker has endorsed this view, and Weisberg has offered a detailed account of how to assess relevant similarities for the purpose of understanding model-world relations (henceforth I will use 'relevant similarities' as shorthand for the relevant similarities which hold between an object and target). I agree with both of them, and think that relevant similarity assessments are the key piece in a framework for evaluating inferences. But they are not the only piece. Hence I see my account as adopting both of their views but also taking a step further toward a complete picture of the considerations which should go into judgments about the external validity of cases of scientific research.

In Parker's discussion of the relationship between experiment and simulation (2009) she rejects what I've called the materiality thesis on grounds which are complementary to the points I made in Chapter 4 (see discussion in Section 4.3). She argues that instead of focusing on material correspondence to evaluate inferences we should focus on relevant similarities. Parker says that "the focus on materiality is somewhat misplaced here [in the context of the materiality thesis], because it is relevant

[30] This is a preliminary account of what I take to be the most important considerations, and not necessarily an exhaustive list. The examples I discuss in this preliminary account focus on responding to the issues that came up in Chapter 4 regarding experiments and simulations. In future work I hope to develop a more nuanced account of how to think about evaluating inferences in these terms in the different kinds of experiments outlined in Section 1.2 and Chapter 3 (laboratory versus field versus natural experiments, and hypothesis-testing versus exploratory experiments), and observational studies.

similarity, not materiality, that ultimately matters when it comes to justifying particular inferences about target systems" (2009, p. 484). Relevance is a function of the question being asked about the target: Depending on what a researcher wants to know about his target, capturing some subset of its features in his object of study will be crucial, and other features might be ignored. The relevant similarities in a given case might be of the sorts that Morgan and Guala called material, formal, or some combination of the two; it depends on what the researcher is after. I completely agree with Parker, but more needs to be said. She does not give a thorough account of how to measure or evaluate relevant similarities, or how we judge them to have been achieved (that is, how we are to rely on them as an alternative basis to material object–target correspondence for judging inferences valid). Picking up where Parker left off, I think we need a more nuanced account of what 'relevant similarity' means and how it figures into experimental design, analysis, and inference.

It would be hard to object to Parker's claim that the key thing scientists need to know in order to justify an inference from an object to a target is "whether the experimental and target systems were actually similar in the ways that are relevant, given the particular question to be answered about the target system" (2009, p. 493). I think it is fair to say that Morgan and Guala both accept this claim in their papers defending the materiality thesis. They just prioritize one particular way to be relevantly similar. While they don't put it in exactly these terms, one way to think about their materiality thesis, which I do not think they would object to, is as follows: Material correspondence is sufficient for capturing the relevant object–target similarity needed to validate an inference. In other words, materiality matters enough that achieving material

correspondence is tantamount to achieving ideal conditions for a valid inference. As I discussed in Chapter 4, this is problematic.

Other people talk about relevant similarity in roughly the sense Parker has in mind, too. This point is often relied on, but seldom articulated in detail: In order to successfully learn about one thing by studying another, the two should be similar in the relevant ways for inferring something about the former from the latter to be credible. For example Samir Okasha, discussing research strategies for understanding the major transitions in evolution by studying natural selection acting at multiple levels, asks: "… is the transition from unicellularity to multicellularity relevantly similar to the transition from solitary insects to eusocial insect colonies? If so, then can the theoretical principles needed to understand the former be extrapolated to the latter and vice versa?" (Okasha 2008, p. 152).

We can begin to flesh out a more developed account of how to think about relevant similarities by returning to two familiar cases from laboratory natural selection experiments discussed in Chapters 2 and 4, punctuated evolution and high mutation rates, and distinguishing them in terms of the relevant similarities which are meant to be grounding the two inferences in question. I will say a bit about what sort of justification there is supposed to be for these two claims, and then make some evaluative remarks about how well this pans out in each case. I said in Chapter 4 that the two cases differ in the degree of material correspondence that is meant to be playing a role in grounding the inference in question. As a starting point for distinguishing them now in terms of the sorts of relevant similarities at play, it is helpful to rely on the division of the experimental

system into three of its key aspects: organisms, their environments, and the governing mechanisms and processes.[31]

In the punctuated evolution case, a *process* was meant to be front and center in grounding the inference, namely, the process of punctuated dynamics in morphological evolution. The environment's constancy was important, but its particular constitution was not. The particular identity (phylogenetic, phenotypic, or otherwise) of the organisms comprising the evolving populations was not a consideration at all. So in this case, weighing the relevance of each of these three aspects (processes, environment, and organisms) matters a lot more than some overall assessment of material correspondence like the one I laid out in Section 4.3.2, which defaults to focusing on only the latter two aspects being alike in the object and target.[32] Whether or not the same dynamics and underlying mechanism are captured in the object and target is key to answering the question of whether or not this is a valid inference, and the choice of experimental system (in this case, twelve strains of *E. coli* in minimal growth medium) can be geared toward isolating features crucial for understanding long-term evolutionary dynamics. For trying to make an inference like the one in the punctuated evolution case, making the experimental system more like the world outside of the laboratory—for example, by introducing environmental variation over space and time or using organisms plucked directly from the wild rather than developed for use as laboratory model organisms—

---

[31] I do not mean to imply that in practice these are three modular systems which can be considered or manipulated independently of each other. They are closely interconnected and manipulations of one often cause changes in the others. I am just separating them conceptually as three potential aspects of comparison between an object and a target.

[32] Or, perhaps, only the organisms. In their discussions of material correspondence Morgan and Guala both focus on the entities or organisms which are the experimental subjects, and do not say much even about the environmental conditions of the experimental system.

would detract from its value by undermining its goal of isolating and capturing the features most salient for studying the high-level evolutionary dynamic in question.

The inference in the high mutation rates case is different. There, the identity of the experimental populations is central to validating the inference from *E. coli* to asexual pathogenic microbes outside of the laboratory. Here is a case where being materially closer to the target matters more, in the sense I discussed in Section 4.3.2, where material correspondence is understood in terms of grain of correspondence at the relevant "material" (that is, chemical, physical, biological…) level of categorization. In this case the relevant categorization is phylogenetic combined with a more general point about reproductive class (being an asexual, non-recombining population). But this point about material closeness mattering holds in this case because close correspondence of physical/biological traits of the experimental subjects and the entities which comprise the target is what needs to be the case for this to be a plausible inference. It has to do with the level of phylogenetic and physiological specificity with which the target is defined in this case—not with the fact that this is an experiment, per se.

We can now return to the question, which I raised and set aside in Chapter 4, of how to actually evaluate the inferences made in these two cases. I mentioned the intuition that the inference in the high mutations rates case seems more plausible than that in the punctuated evolution case. If we analyzed these two cases from the perspective of the material correspondence view, we would get the following upshot. The high mutation rates case is in a good position with respect to external validity, since its experimental subjects are "made of the same stuff" as the target organisms in precisely the same way as in Morgan's examples from economics. The punctuated evolution case would not get such a clear stamp of approval, but this is because, as I discussed in Section 4.3.2, it gets messy

trying to even make sense of what material object–target correspondence would look like. The point of the inference in that case is not to delineate a target in terms of any particulars of its material constitution; in a material sense, the target is not concretely defined. It is an arbitrary set of populations, traces of whose evolutionary history are left in the fossil record; the target in this case is about the macroevolutionary trend of punctuated equilibrium. The aim in that case was to capture what is going on with a certain kind of evolutionary dynamic and its historical causes. Thus, the material correspondence view seems to more or less line up with our intuitions about these two cases, but for the wrong reason. It fails to point to a valid inference in the punctuated evolution case, but only because it fails to offer a coherent analysis of what is going on in that case.

If we think about these two cases in terms of capturing relevant similarities, rather than material correspondence, we can get a different analysis. The high mutation rates case (see Section 2.1.1) looks plausible because the relevant similarities are plausibly captured between the asexual, non-recombining populations of evolving microbes in the experimental system and those designated as the target. The claim being made in that case has the following form:

1. We observe the evolution of high mutation rates in the lab populations.
2. We posit a mechanism for how this might occur, mutator hitchhiking.
3. The same mechanism might explain the evolution of high mutation rates observed in similar populations in nature, with phylogenetic classification identical to or close to that of the lab populations, sharing the key feature that makes mutator hitchhiking work: lack of recombination.

This looks like a good inference from object to target for roughly the same reason as discussed in the preceding paragraph, but rather than just talking about material correspondence we can put it in terms of relevant similarity—though, if you like, it is still "material correspondence" (in this case, cashed out as corresponding classification as members of overlapping or nearby phylogenetic classes and as asexual populations) doing the work. The point is, the way the claim is made, researchers give solid reasons for thinking that the features relevant to explaining the mechanism in question hold in both the lab populations (object) and the populations in nature (target). Both sets of populations are relevantly similar with respect to those key characterizing features.

The punctuated evolution case (see Section 2.1.2) looks questionable for a more satisfying reason than on the material correspondence analysis discussed above, where it was just unclear how to understand the target's "material" status and thus how to even go about evaluating material correspondence. The inference from the lab populations to the fossil record is dubious because it is not clear that the laboratory populations' short-term punctuated evolutionary dynamic actually has to do with the massively longer-term proposed evolutionary dynamic of punctuated equilibrium in nature. The inference there has the following form:

1. We observe punctuated morphological evolution over 3,000 generations in the lab populations (the cells increase in size over a short number of generations, stay the same size for many generations, and then increase rapidly again).
2. We posit a mechanism for how this might occur, rare beneficial mutations rapidly sweeping to fixation.
3. The same mechanism might explain punctuated equilibrium observed in the fossil record, where macroevolutionary trends appear to show long periods of

> stasis punctuated by short periods of rapid change. But—the authors do not note this but I am noting it here—the way Gould and Eldredge (Eldredge & Gould 1972; Gould & Eldredge 1977) talked about punctuated equilibrium in their classic discussion, in addition to the massive difference in scale, the periods they focus on are concentrated periods of rapid speciation events, not evolution of single morphological traits.

In this case, we have to focus on the evolutionary dynamics at play, and functional features of the environment, to be in a position to evaluate the inference from object to target. Material correspondence is not the point. And there are not convincing reasons given here for why the laboratory populations should be taken as relevantly similar to a set of fossilized organisms (and the long-gone lineages they represent) taken as evidence of punctuated equilibrium, for the purpose of establishing the claim the authors make. The problem is that the authors use observations about relatively very short-term dynamics in single lineages in the laboratory to propose a mechanism for hugely longer-term dynamics in nature spanning speciation events, without doing the work to show why the object is a valid stand-in for the target. There is a major scale difference between the laboratory dynamics and target dynamics in question, combined with an undiscussed difference between the observation of a single morphological trait (cell size) evolving in a few leaps, versus a complex of observations of changes over speciation events at issue in discussions of punctuated equilibrium.

The view that material correspondence is all that matters does not leave enough room for tailoring what matters in an inference claim to the kind of scientific question being asked or answered. Thinking in terms of relevant similarities addresses this: Depending on the kind of inference being made, different aspects of the relationship

between the object of study and the outside world will be more salient and crucial to capture. External validity need not be proportional to degree of material object-target correspondence. Instead, external validity assessments should account for what is relevant about correspondence to the outside world not only of the entities which make up the object of study, but also a host of other potential features of the governing processes and environment (particularly in experiments).

Recall the quotation from Guala on p. 75 above, where he says that "The trick [to generating external validity] is to make sure that the target and the experimental system are similar in most relevant respects… Experimenters make sure that this is the case by using materials that resemble as closely as possible those of which the parts of the target system are made." Guala starts out on the right track here, but he makes the mistake of conflating achieving material correspondence in an experiment with capturing what is most relevantly similar about the experimental system and the target. Material correspondence is *one possible route* to achieving external validity. Depending on the context, it might or not might be the best route, or even an appropriate route. I am not saying that material correspondence never matters. My point is just that it is wrong to think that there is something special about experimentation such that material correspondence is the crucial, or even central, feature grounding valid extrapolation claims. Sometimes it is; sometimes it is not. Relevant similarities between objects and targets should be the basis of extrapolation claims, and these can be similarities of many different sorts—physiological, functional, mechanical, material, phylogenetic, mathematical, "formal," etc.—depending on the scientific question at hand.

One of the most complete accounts of the notion of similarity is from Weisberg (Weisberg 2013). Weisberg focuses on model-target correspondence in modeling and

simulations, but his account could easily be adopted to apply to any case where we are assessing object-target correspondence. Weisberg proposes a formal account for understanding the similarity between a model and its target based on the idea of *weighted feature matching*. This involves compiling a list of the salient features of the model and the salient features of the target, and calculating which features they share and which features they fail to share. Features are divided into attributes (properties and patterns) and mechanisms (the underlying processes which generate attributes). The formal metric for assessing similarity on Weisberg's account involves calculating the intersections of attributes and of mechanisms shared between model and target, as well as the difference between the model's and target's respective attributes and mechanisms. Each of the resulting six terms is then assigned a weight, based on the relative importance of each member in the feature set to the researcher's modeling goals. Overall similarity is the ratio of weighted shared features to weighted features which are not shared. For further details of the account and several examples of its application, see (Weisberg 2013, chapter 7).

I consider this an account of *relevant* similarity in the sense I am talking about here because it incorporates context and what matters to the researcher. Weisberg says:

> A model is similar to its target… when it shares certain highly valued features, doesn't have many highly valued features missing, and then the target doesn't have many significant features that the model lacks. Relevant features are identified in a natural or formal language and their importance is weighted relative to the goals of the scientific community. (2013, pp. 144–145)

The relevance comes in in the use of (only) highly valued, significant features, and in their weighting. This is an account of how we should think about relevant similarities: roughly, the feature matching captures the similarity, and the weighting captures the relevance. Scientists thinking about object-target correspondence on Weisberg's account have to

make decisions about which features of their object and target are most salient for establishing external validity; not all similarities count equally.

Weisberg illustrates his account with examples including the San Francisco Bay model and Schelling's segregation model. The former is a huge (1.5 acre) scale model of the San Francisco Bay with hydraulic pumps simulating tides and currents, used to study the effects on the bay of actions such as constructing potential dams. Schelling's segregation model is a computational model showing how racial segregation can occur in the absence of explicitly racist attitudes. It consists of a spatial grid populated by agents of different types, each of whom has a slight preference that their neighbors be of the same type as them; the model can be applied to thinking about the mechanisms of segregation in particular cities like Philadelphia (see further discussion of both cases in Weisberg 2013). Weisberg explains how we think in terms of weighted feature matching to assess the extent to which object (model) and target share the features researchers care about.

We can think of the resulting calculations as good starting points for evaluating the validity of inferences from objects to targets: a higher score would entail a greater license to place confidence in such inferences because it points to greater reason to be confident that the object has captured the features of the target that matter. We can successfully carry out these sorts of similarity assessments only when we know quite a bit about our target's attributes and mechanisms. This works well in the examples Weisberg uses to illustrate his account. In the case of the San Francisco Bay model, we know a lot about the bay. In the case of the Schelling model being used to model racial segregation in Philadelphia, we know a lot about Philadelphia. We do not know everything about either target; if we did, we wouldn't be asking questions about them and trying to learn more. But we know enough about their relevant attributes and mechanisms to generate a

satisfactory assessment of their similarity to those of the object (model). For instance, we can be confident that the simple shared utility functions used in the model (such as "I want at least 30% of my neighbors to be like me") are not strictly identical to those held by the actual citizens of Philadelphia, and we know that the rigid grid structure of the model's spatial layout does not strictly match the irregular grid structure of the actual city of Philadelphia but still does a fine job of approximating it in various ways (Weisberg 2013, p. 148).

In cases where we know much less about the target than we know about the San Francisco Bay or the grid and population structures of US cities, we would be in a worse position to assess relevant similarities because we might not even know enough to calculate exactly which attributes and mechanisms are shared, or fail to be shared. In many cases, scientists know a lot less about their target's attributes and mechanisms than we know about the targets in these examples. To be fair to Weisberg, he intends weighted feature matching calculations to represent the relation that is supposed to hold when relevant similarities are achieved. He does not say that a modeler will always be in a good position to perform the relevant calculations about her model and target, for the sort of reason I'm raising here (she might know little about her target—or about her object). My point is not to criticize his account, but just to underline that many cases are far enough from the ideal that weighted feature matching calculations will be difficult or impossible to carry out. For example, this will often be the case in research areas whose targets are entities and phenomena in the deep past, like research on long-extinct species or the origin of life of Earth. In such cases, we cannot rely only on successful feature-matching calculations to give us confidence that our object and target are sufficiently relevantly similar, and thus to place confidence in inferences from the former to the latter.

I think Weisberg's account is a good way to think about assessing relevant similarities, and I am not proposing an alternative way to assess them. Rather, I'm underlining that we cannot always carry out clear relevant similarity assessments. For the present purposes—an account of how to evaluate inferences from objects to targets—we should focus centrally on relevant similarities, but also take a step back and have some framework in place for judging inferences even when we cannot carry out these assessments. Sometimes a scientist's epistemic relationship to her target, or even to her object, is such that she is not in a good position to say anything too definitive about what similarity relationship holds between them. In particular, I'll making some preliminary remarks about some further considerations that matter: how much a scientist knows about her target and her object, and the closely related issue of the value of control versus realism. When she knows less about her object or target or especially both, realism matters more, and we often have reason to place greater confidence in an experiment than in a computer simulation. When she knows a lot about both, realism can sometimes be sacrificed for control, and we can sometimes have reason to place greater confidence in a simulation than we would in an experiment.

First, regarding the cases when we know less about the target, when realism matters more: We can aim for relevant similarities even when we are not in a position to undertake a clear feature-by-feature assessment as in Weisberg's account. The issue of identifying and assessing particular features which make an object and target relevantly similar is not the same as having reasons to think that the object might be relevantly similar to the target. Usually these go hand in hand, but sometimes we can have the latter without a clear idea of the former. This seems to be the case especially in biology, where we often rely on phylogenetic closeness to justify studying one organism as a stand-in for

another, when the former is a model organism we know a lot about and the latter is an organism in nature we know much less about. The reason to think the two might be relevantly similar can involve knowing, for example, that the species evolved by natural selection in similar environments, or that they share key homologous structures of interest. Even if we cannot list all of the features they share, that is a good reason to think that they are similar in ways that matter. On this sort of point, Maclaurin and Sterelny write:

> To say that two organisms are members of the same taxonomic group is to say that they are importantly similar, and, depending on the taxonomic system employed, to license inferences based on those similarities. Importantly, they will be similar with respect to features whose existence or importance we are yet to discover… The enormous scientific effort expended on understanding the developmental biology of just a few model organisms (a fruit fly, a nematode worm, a mouse, a fish) is based on this intellectual strategy. (Maclaurin & Sterelny 2008, p. 10).

To reiterate an important point I made in Chapter 4: When we have little background knowledge of our target, a physical sample of our target or a close approximation is often the best starting point. Experiments often have epistemic privilege in these contexts where we know little about our target.

When we know more about the target, control can matter more than realism. In the contexts of some research areas, we know enough to build reliable simulations precisely because we have enough information from the world already. The molecular bond example discussed at the end of Chapter 4 (see page 105) is such a context: We know enough about the way individual atoms fit into molecules, and the effects of atomic changes on molecular structure, to design reliable simulations. These are the sorts of cases where not only can we place confidence in the simulation results, but it seems right to say that simulations have privilege over experiments. Carrying out the relevant experimental manipulation (that is, physically swapping out the atoms of phosphorus in an actual DNA

molecule for atoms of arsenic) would be extremely difficult, and would arguably involve steps of uncertainty and need for confirmation (has the intended intervention actually been successfully achieved?) which would give us reasons to have decreased confidence in the experimental result compared to the equivalent well-designed simulation.

Other cases present even more strikingly how sometimes, however much we know or fail to know about the object or target, it makes for better inferences to sacrifice realism for control for pragmatic reasons. In these cases, simulations have privilege over experiments in a different sense: They are for practical reasons the best we can do. For example, there are situations where studying a physical approximation of the target would be unfeasible, such as large-scale climate studies (see Parker 2009). In these cases, studying a physical sample or a close physical analogue of the climate of the entire planet does not make sense as a feasible experimental program. This is a point about hugely macro-scale targets of inquiry; the same sort of point also holds at the extreme micro-level, for examine in certain cases in nanoscale physics. While it is not strictly impossible to undertake the experimental manipulations in question, it would be prohibitively difficult (and would perhaps require technology we have not yet developed), but more importantly we have enough theoretical knowledge to construct reliable computer models (see discussion in Lenhard 2006).

Here is another example of a case where control seems to matter more than realism, in this case because a certain kind of question is being asked about general dynamics rather than particular kinds of organisms or entities in the world. This example involves trying to figure out the relative importance of different evolutionary processes in populations with high mutation rates. In a recent paper Keller and colleagues (2012) discuss computer simulations aimed at understanding which processes are responsible

when null models of mutation-selection balance fail to predict a population's fitness equilibrium. They explain their choice of object of study as follows: "we do not pretend that our model captures any biological system. The property that is most appealing is a fitness landscape in which many different biological properties can evolve" (p. 2308). In a case like this, control of certain high-level features is paramount, and there are a number of reasons to think that studying a material system, like laboratory populations of organisms, would put researchers in an epistemically worse situation with respect to answering the particular question at hand, namely: When the null models fail to predict fitness equilibrium, what sorts of other evolutionary processes might be responsible? This is because they would be sacrificing much-needed control for arguably unneeded realism (for example, it would be hugely more difficult to identify and measure the fittest genotypes in the population.) This is where the kind of intuition underlying the materiality thesis comes in: If we are asking a scientific question that relies particularly on physical, physiological, or phylogenetic object–target correspondence, experiments are the best route to valid inferences. It is the conditional that is key here: Not all scientific questions rely on such correspondence to achieve valid inferences about the target in question; in fact, some explicitly have goals that conflict with such correspondence.

The aim of this chapter was to outline the kinds of considerations that should go into a framework for evaluating inferences from objects to targets which allows for context-sensitive judgments about a range of factors including, but not limited to, the identity of the object in question (computer model or physical system). This was to shift away from the problematic basis of basing such judgments purely on methodological categories like experiment or simulation. I talked about the importance of three interrelated considerations for such an account: relevant similarities, how much we know

about the object and target in question, and how we weigh the importance of control versus realism. There's a lot more to be said about each of these three considerations; this is a topic for future work. The main point I want to establish here is that these are the considerations that should go into such an account.

In closing, from the discussion in this dissertation a number of lessons follow for how we think about experiments and their role in scientific inquiry. First, the categorization scheme differentiating hypothesis-testing and exploratory experiments is not exhaustive; much work remains to be done to clarify the middle ground. Second, attempts to draw clear methodological and epistemic lines dividing experiments and simulations break down, especially in the case of experiments versus what I called broad-sense simulations. Third, philosophers, scientists, funding agencies, and others evaluating particular cases of scientific research should give up the common practice of relying the experiment/simulation line as a basis for in-principle judgments about the value of those cases. Instead, we should focus on questions about the relationship between objects and targets in particular contexts, what the researcher knows about both, and what sort of question she is trying to answer by studying her object. In arguing for these points I have argued neither that methodological categories are meaningless,[33] nor that experiments never have epistemic privilege over simulations. As the preceding discussion in this chapter especially shows, there are important differences between studying physical experimental systems and studying computer models. But—this is the key point—our judgments of the value of these sorts of methodological choices for backing up good scientific inferences must be sensitive to the context of the research question, and research area background knowledge, at hand. The right way to think about evaluating inferences is

[33] Except in the case of experiments versus "physical simulations"—that, I think, is a meaningless distinction.

in terms of relevant similarity assessments (when possible), combined with thinking about how much a researcher knows about her object of study and target of inquiry, and the relative importance of control versus realism to generating a good answer to the scientific question at hand.

Bibliography

Abrams, M. (2012). Measured, modeled, and causal conceptions of fitness. *Frontiers in Genetics*, 3 (196).

Atwood, K., Schneider, L., & Ryan, F. (1951). Periodic Selection in Escherichia Coli. *Proceedings of the National Academy of Sciences of the United States of America*, 37(3), 146.

Barberousse, A., Franceschelli, S., & Imbert, C. (2008). Computer simulations as experiments. *Synthese*, 169(3), 557–574.

Barnett, S. A., & Dickson, R. G. (1984). Changes among wild House mice (*Mus musculus*) bred for ten generations in a cold environment, and their evolutionary implications. *Journal of Zoology*, 203(2), 163–180.

Bataillon, T., Joyce, P., & Sniegowski, P. (2013). As it happens: current directions in experimental evolution. *Biology Letters*, 9(1), 20120945.

Bateman, M. S. (2012). *The dynamics of inquiry in cognitive neuroscience*. Dissertations available from ProQuest. Paper AAI3550754. http://repository.upenn.edu/dissertations/AAI3550754 (accessed March 2015).

Beatty, J. (2006). Replaying life's tape. *The Journal of Philosophy*, 103(7), 336–362.

Bedau, M. A. (2008). Weak Emergence. *Noûs*, 31, 375–399.

Bell, G. (2013). Evolutionary rescue of a green alga kept in the dark. *Biology Letters*, 9(1), 20120823.

Blount, Z. D., Barrick, J. E., Davidson, C. J., & Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417), 513–518.

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23), 7899–7906.

Bono, L. M., Gensel, C. L., Pfennig, D. W., & Burch, C. L. (2013). Competition and the origins of novelty: experimental evolution of niche-width expansion in a virus. *Biology Letters*, 9(1), 20120616.

Brandon, R. N. (1994). Theory and experiment in evolutionary biology. *Synthese*, 99(1), 59–73.

Burian, R. M. (2007). On microRNA and the need for exploratory experimentation in post-genomic molecular biology. *History and Philosophy of the Life Sciences*, 29(3), 285–311.

Calcott, B., & Sterelny, K. (Eds.). (2011). *The major transitions in evolution revisited*. The MIT Press.

Colegrave, N., Kaltz, O., & Bell, G. (2002). The ecology and genetics of fitness in chlamydomans. VIII. The dynamics of adaption to novel environments after a single episode of sex. *Evolution*, 56(1), 14–21.

Dallinger, W. H. (1887). The president's address. *Journal of the Royal Microscopical Society*, 7(1), 185–199.

Darden, L., & Cain, J. A. (1989). Selection type theories. *Philosophy of Science*, 56(1), 106–129.

Darwin, C. (2009). *The annotated Origin: A facsimile of the first edition of On the Origin of Species*. Cambridge: Belknap Press.

De Monte, S., & Rainey, P. B. (2014). Nascent multicellular life and the emergence of individuality. *Journal of Biosciences*, 39(2), 237–248.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, 88(1), 27–51.

Denning, E. J., & MacKerell, A. D. J. (2011). Impact of arsenic/phosphorus substitution on the intrinsic conformational properties of the phosphodiester backbone of DNA investigated using ab inition quantum mechanical calculations. *Journal of the American Chemical Society*, 133(15), 5770–5772.

Desjardins, E. (2011). Reflections on path dependence and irreversibility: Lessons from evolutionary biology. *Philosophy of Science*, 78(5), 724–738.

Diamond, J. M. (1983). Laboratory, field and natural experiments. *Nature*, 304(18), 586–587.

Drake, J. (1993). General antimutators are improbable. *Journal of Molecular Biology*, 229(1), 8–13.

Eigen, M. (1971). Self organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10), 465–523.

Eigen, M. (2002). Error catastrophe and antiviral strategy. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21), 13374.

Eldredge, N., & Gould, S. J. (1972). Punctuated equilibria: An alternative to phyletic gradualism. In T. J. M. Schopf (Ed.), *Models in paleobiology* (pp. 82–115). San Francisco: Freeman, Cooper and Company.

Elena, S. F., Cooper, V. S., & Lenski, R. E. (1996). Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272(5269), 1802.

Elliott, K. (2007). Varieties of exploratory experimentation in nanotoxicology. *History and Philosophy of the Life Sciences*, 29(3), 1–21.

Epstein, J. M., & Axtell, R. L. (1996). *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press.

Fijalkowska, I. J., & Schaaper, R. M. (1993). Antimutator mutations in the alpha subunit of *Escherichia coli* DNA polymerase III: Identification of the responsible mutations and alignment with other DNA polymerases. *Genetics*, 134(4), 1039–1044.

Forber, P. (2009). Spandrels and a pervasive problem of evidence. *Biology and Philosophy*, 24(2), 247–266.

Forde, S. E., & Jessup, C. M. (2009). Understanding evolution through the phages. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 391–418). University of Chicago Press.

Franklin, A. (1989). *The neglect of experiment*. Cambridge University Press.

Franklin, A. (1990). *Experiment, right or wrong*. Cambridge University Press.

Franklin, L. (2005). Exploratory experiments. *Philosophy of Science*, 72(5), 888–899.

Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria: An International Journal for Theory, History and Foundations of Science*, 21(1), 49–65.

Frigg, R., & Hartmann, S. (2012). Models in science. *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), E. N. Zalta (ed.), http://plato.stanford.edu/archives/fall2012/entries/models-science/ (accessed March 2015).

Fry, I. (2000). *The emergence of life on Earth: A historical and scientific overview*. Rutgers University Press.

Fukushima, M., Kakinuma, K., & Kawaguchi, R. (2002). Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the gyrB gene sequence. *Journal of Clinical Microbiology,* 40(8), 2779–2785.

Futuyma, D. J., & Bennett, A. F. (2009). The importance of experimental studies in evolutionary biology. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 15–30). University of Chicago Press.

Galison, P. (1987). *How Experiments End*. University of Chicago Press.

Garland, T. J., & Rose, M. R. (2009). *Experimental evolution: Concepts, methods, and applications of selection experiments*. University of California Press.

Gentile, C. F. (2012). *The evolution of a high mutation rate and declining fitness in asexual populations*. Dissertations available from ProQuest. Paper AAI3509062. http://repository.upenn.edu/dissertations/AAI3509062 (accessed March 2015).

Gentile, C. F., Shaver, A., & Parke, E. C. (in preparation). General antimutators are rare in *E. coli*: Conditional antimutator effects of *dnaE911*.

Gentile, C. F., Yu, S.-C., Serrano, S. A., Gerrish, P. J., & Sniegowski, P. D. (2011). Competition between high- and higher-mutating strains of Escherichia coli. *Biology Letters*, 7, 422-424.

Gerstein, A. C. (2013). Mutational effects depend on ploidy level: All else is not equal. *Biology Letters*, 9(1), 20120614.

Gifford, D. R., de Visser, J. A. G. M., & Wahl, L. M. (2013). Model and test in a fungus of the probability that beneficial mutations survive drift. *Biology Letters*, *9*(1), 20120310.

Glass, D. J. (2014). NIH grants: Focus on questions, not hypotheses. *Nature*, 507, 306.

Gould, S. J. (1989). *Wonderful life: The Burgess Shale and the nature of history*. W. W. Norton & Company.

Gould, S. J., & Eldredge, N. (1977). Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, *3*(2), 115–151.

Grant, P. R., & Grant, B. R. (2002). Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, 296(5568), 707–711.

Grimm, V., & Railsback, S. F. (2013). *Individual-based modeling and ecology*. Princeton University Press.

Guala, F. (2002). Models, simulations, and experiments. In L. Magnani and N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 59–74). Kluwer.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge University Press.

Harré, R. (2003). The materiality of instruments in a metaphysics for experiments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 19–38). Pittsburgh: University of Pittsburgh Press.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, 31(11), 4178–4187.

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice Hall.

Holmes, F. L. (2001). *Meselson, Stahl, and the replication of DNA*. Yale University Press.

Huey, R. B., & Rosenzweig, F. (2009). Laboratory evolution meets Catch-22. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 671–701). University of Chicago Press.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

Irschick, D. J., & Reznick, D. (2009). Field experiments, introductions, and experimental evolution. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 173–194). University of California Press.

Jessup, C. M., Kassen, R., Forde, S. E., Kerr, B., Buckling, A., Rainey, P. B., & Bohannan, B. J. M. (2004). Big questions, small worlds: Microbial model systems in ecology. *Trends in Ecology & Evolution*, 19(4), 189–197.

Jorde, L. B., Bamshad, M., & Rogers, A. R. (1998). Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays*, 20(2), 126–136.

Keller, E. F. (2003). *Making sense of life*. Harvard University Press.

Keller, T. E., Wilke, C. O., & Bull, J. J. (2012). Interactions between evolutionary processes at high mutation rates. *Evolution*, 66(7), 2303–2314.

Kerr, B., Riley, M. A., Feldman, M. W., & Bohannan, B. J. M. (2002). Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, 418(6894), 20120569.

Khuyagbaatar, J., Yakushev, A., Düllmann, C. E., Ackermann, D., Andersson, L.–L., Asai, M. et al. (2014). Ca 48+ Bk 249 Fusion Reaction Leading to Element Z= 117: Long-Lived α-Decaying Db 270 and Discovery of Lr 266. *Physical Review Letters*, 112, 172501.

Landman, U., Luedtke, W. D., Burnham, N. A., & Colton, R. J. (1990). Atomistic mechanisms and dynamics of adhesion, nanoindentation, and fracture. *Science*, 248(4954), 454–461.

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., et al. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39), 15716–15721.

Lenhard, J. (2006). Surprised by a nanowire: Simulation, control, and understanding. *Philosophy of Science*, 73(5), 605–616.

Lenhard, J. (2007). Computer simulation: The cooperation between experimenting and modeling*. *Philosophy of Science*, 74(2), 176–194.

Lenski, R. E. (2011). Evolution in action: A 50,000-generation salute to Charles Darwin. *Microbe*, 6(1), 30–33.

Lenski, R. E. (2015). The E. coli long-term experimental evolution project site. http://myxo.css.msu.edu/ecoli (accessed February 2015).

Lenski, R. E., & Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America*, 91(15), 6808.

Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*, 136(6), 1315–1341.

Levy, A., & Adrian, C. (forthcoming). Model organisms aren't (theoretical) models. *British Journal for the Philosophy of Science*.

Love, A. C., & Travisano, M. (2013). Microbes modeling ontogeny. *Biology and Philosophy*, 28(2), 161–188.

Maclaurin, J., & Sterelny, K. (2008). *What Is biodiversity?* University of Chicago Press.

MacLeod, M., & Nersessian, N. J. (2013). Building simulations from the ground up: Modeling and theory in systems biology. *Philosophy of Science*, 80(4), 533–556.

Matsuba, C., Ostrow, D. G., Salomon, M. P., Tolani, A., & Baer, C. F. (2013). Temperature, stress and spontaneous mutation in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *Biology Letters*, 9(1), 20120334.

Maynard Smith, J. (1978). *The evolution of sex*. Cambridge University Press.

Maynard Smith, J., & Szathmary, E. (1997). *The major transitions in evolution*. Oxford University Press.

McClintock, B. (1951). Chromosome organization and genic expression. *Cold Spring Harbor Symposia on Quantitative Biology*, 16(0), 13–47.

Morgan, M. S. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, 12(2), 317–329.

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies*, 143(1), 33–57.

Muller, H. (1964). The relation of recombination to mutational advance. *Mutation research/fundamental and molecular mechanisms of mutagenesis*, 1(1), 2–9.

Novick, A., & Szilard, L. (1950). Experiments with the chemostat on spontaneous mutations of bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 36(12), 708.

O'Malley, M. A. (2007). Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences*, 29(3), 335–358.

Ofria, C., & Wilke, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, 10(2), 191–229.

Okasha, S. (2008). The units and levels of selection. In S. Sarkar & A. Plutynski (Eds.), *A companion to the philosophy of biology*, Chapter 8.

Okasha, S. (2011). Experiment, observation and the confirmation of laws. *Analysis*, 71(2), 222–232.

Parke, E. C. (2014a). Flies from meat and wasps from trees: Reevaluating Francesco Redi's spontaneous generation experiments. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 45, 34–42.

Parke, E. C. (2014b). Experiments, simulations, and epistemic privilege. *Philosophy of Science*, 81(4), 516–536.

Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. *Synthese*, 169(3), 483–496.

Peck, S. L. (2004). Simulation as experiment: A philosophical reassessment for biological modeling. *Trends in Ecology & Evolution*, 19(10), 530–534.

Peschard, I. (2012). Is simulation an epistemic substitute for experimentation? In S. Vaienti (Ed.), *Simulations and networks* (pp. 1–17). Paris: Hermann.

Popper, K. (1934). *The Logic of Scientific Discovery*. New York.

Quinones, A., & Piechocki, R. (1985). Isolation and characterization of *Escherichia coli* antimutators. *MGG Molecular & General Genetics*, 201(2), 315–322.

Radder, H. (Ed.). (2003). *The philosophy of scientific experimentation*. University of Pittsburgh Press.

Railsback, S. F., & Grimm, V. (2011). *Agent-based and individual-based modeling: A practical introduction*. Princeton University Press.

Rainey, P. B., & Rainey, K. (2003). Evolution of cooperation and conflict in experimental bacterial populations. *Nature*, 425(6953), 72–74.

Ratcliff, W. C., Denison, R. F., Borrello, M., & Travisano, M. (2012). Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 109(5), 1595–1600.

Raynes, Y., Gazzara, M. R., & Sniegowski, P. D. (2011). Mutator dynamics in sexual and asexual experimental populations of yeast. *BMC Evolutionary Biology*, 11(1), 158.

Redi, F. (1668). *Esperienze intorno alla generazione degl'insetti*. Original text at http://www.liberliber.it (accessed February 2011).

Roff, D. A., & Fairbarin, D. J. (2009). Modeling experimental evolution using individual-based, variance-components models. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 31–64). University of California Press.

Rose, M. R., & Garland, T. J. (2009). Darwin's other mistake. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 3–13). University of California Press.

Simoes, P., Santos, J., & Matos, M. (2009). Experimental evolutionary domestication. In T. J. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 89–110). University of Chicago Press.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature*, 387(6634), 703–705.

Sprouffske, K., Merlo, L. M. F., Gerrish, P. J., Maley, C. C., & Sniegowski, P. D. (2012). Cancer in Light of Experimental Evolution. *Current Biology*, 22(17), R762–R771.

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, 64, S65–S74.

Travisano, M., Mongold, J. A., Bennett, A. F., & Lenski, R. E. (1995). Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194), 87-90.

Turner, P. E., McBridge, R. C., & Zeyl, C. W. (2009). Sexual exploits in experimental evolution. In T. Garland & M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 479–521). University of Chicago Press.

Turner, P. E., Souza, V., & Lenski, R. E. (1996). Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology*, 77(7), 2119–2129.

Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

Waters, C. K. (2007). The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. *History and Philosophy of the Life Sciences*, 29(3), 1–9.

Weber, M. (2004). *Philosophy of Experimental Biology*. Cambridge University Press.

Weber, M. (2014). Experiment in biology. *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), E. N. Zalta (Ed.), http://plato.stanford.edu/archives/win2014/entries/biology-experiment/ (accessed March 2015).

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Weisstein, E. W. (2013). Game of Life. From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/GameofLife.html (accessed March 2015).

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, 70(1), 105–125.

Winsberg, E. (2009). A tale of two methods. *Synthese*, 169(3), 575–592.

Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.

Woodward, J. (2003). Experimentation, causal inference and instrumental realism. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 87–118). Pittsburgh: University of Pittsburgh Press.

Woodward, J. (2013). Causation and manipulability. *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), E. N. Zalta (Ed.), http://plato.stanford.edu/archives/win2013/entries/causation-mani/ (accessed March 2015).

Zera, A. J., & Harshman, L. G. (2009). Laboratory selection studies of life history physiology in insects. In T. J. Garland and M. R. Rose (Eds.), *Experimental evolution: Concepts, methods, and applications of selection experiments* (pp. 217–262). University of Chicago Press.

Zeyl, C. (2000). Budding yeast as a model organism for population genetics. *Yeast*, 16(8), 773–784.